Improved Prediction of Perceived Fluency of Japanese English using Quantity of Phonation and Quality of Pronunciation [*]

☆ Yang SHEN, Ayano YASUKAGAWA, Daisuke SAITO, Nobuaki MINEMATSU (UTokyo),
Kazuya SAITO (UCL)

1 Introduction

To support people to learn a new language, various types of technical aids have been examined [1, 2] and realized as commercial products or services [3, 4]. This paper presents research results of a joint project between UTokyo and UCL, where teachers asked engineers to automatize their fluency scoring strategy. In this study, prediction is conducted with Elastic Net regression with speech features. Experimental results demonstrate that posteriorgram with multiple granularities is effective for prediction and a high correlation of 0.925 is obtained between machine scores and the scores of perceived fluency averaged over 10 native raters. This value is higher than the average of inter-rater correlations of 0.873.

2 Related works

2.1 Picture description corpus with fluency rating [5]

90 native Japanese studentsas well as 10 native speakers participated in data collection. The task is picture description, where three independent photos were presented with three keywords per photo to the participants, as shown in Figure 1. They were asked to describe the pictures orally using the keywords. Their utterances were recorded with 16 bits and 44.1 kHz as sampling frequency.

10 native raters, who did not participate in data collection, were recruited for manual fluency assessment of the 100 utterances. They are native speakers, but not teachers or researchers of language education. The score varied from 1 (=least fluent) to 9 (=extremely fluent). Before rating, the definition of fluency in [5] was explained to the raters, who showed a high consensus on that definition.

Each rater assigned 100 scores to the 100 utterances. Correlations are calculated between every pair of the raters. The minimum, average, and maximum of one-to-one correlations are 0.677, 0.786, and 0.897. Correlations are also calculated for each rater to the averaged scores of the other nine raters. The minimum, average, and maximum of one-to-others correlations were 0.798, 0.873, and 0.910. These are used as reference when assessing the performance of automatic prediction of fluency.

2.2 Manual extraction of features related to fluency [5]

The following features were manually extracted with Praat [6], 1) the number of breakdowns, (un)filled pauses, per unit time, 2) speaking rate, the number of syllables per unit time, 3) the number of repairs per unit time. The number of breakdowns were counted separately for two cases, within and between clauses. Repairs can also be divided into repetitions and self-corrections. The five features were expected to affect raters' judgements through an extensive review of the related literature.

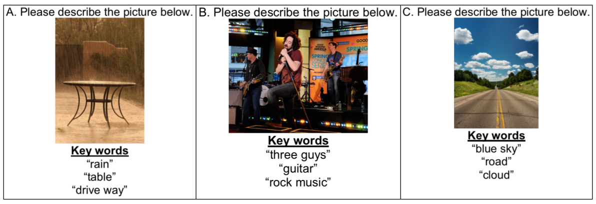
A. Please describe the picture below.
Key words
"rain"
"table"
"drive way"
B. Please describe the picture below.
Key words
"three guys"
"guitar"
"rock music"
C. Please describe the picture below.
Key words
"blue sky"
"road"
"cloud"

Fig. 1 Three photos used for data collection

We regard the above features as related to quantity of phonation, per unit time and applied Elastic Net regression to predict the fluency scores, averaged over the 10 raters. 5-fold cross-validation showed that the predicted scores had a correlation of 0.788 to the human scores, which is comparable to the average of one-to-one correlations. This value can be used as reference when assessing the performance of automatic prediction of fluency.

2.3 Clustering of phonemic classes using posteriors [7]

Besides features related to quantity of phonation, those related to quality of pronunciation are also examined. For this end, all the utterances are converted to posteriorgrams. Posteriorgrams generally use a set of phoneme classes, the number of which is several thousands. They can be viewed as finely-defined context-dependent phonemes, but they may be too fine to be used for assessment. We reduce the number of classes using bottom-up clustering with Ward's method [8] , which requires the distance matrix between any two classes. The Bhattacharyya distance between two classes a and b is re-written using class posterior through Bayes' theorem [7] as

$$\begin{aligned}BD(a,b) &= -\ln\int\sqrt{p(\boldsymbol{x}|a)p(\boldsymbol{x}|b)}d\boldsymbol{x}\\ &= -\ln\int\sqrt{\frac{p(a|\boldsymbol{x})p(\boldsymbol{x})}{p(a)}\frac{p(b|\boldsymbol{x})p(\boldsymbol{x})}{p(b)}}d\boldsymbol{x}\\ &= -\ln\int p(\boldsymbol{x})\sqrt{p(a|\boldsymbol{x})p(b|\boldsymbol{x})}d\boldsymbol{x}+\frac{1}{2}\ln p(a)+\\ &\quad \frac{1}{2}\ln p(b).\end{aligned}$$

[*] 発声の量と発音の質に関する特徴を用いた日本人英語の主観流暢さ予測の改善
瀋陽, 安ヵ川彩乃, 齋藤大輔, 峯松信明（東大）, 斉藤一弥（UCL）

$p(\boldsymbol{x})$ is a prior probability for $\boldsymbol{x}$, which can be calculated using the universal background model. $p(a|\boldsymbol{x})$ and $p(b|\boldsymbol{x})$ are class posteriors, which are outputs from DNN-based acoustic models to input vector $\boldsymbol{x}$. $p(a)$ and $p(b)$ are prior probabilities for the two classes, which can be obtained as normalized frequency from the training corpus. Once DNN models are trained, any speech sample can be converted to its posteriorgram, which is a sequence of vectors comprised of probabilities of phoneme classes. With the above formulation, a given posteriorgram can be reduced into a smaller dimension of classes. In the current study, the baseline number of classes is 2,000 and n-class posteriorgrams can be calculated for any n ($2 \leq n \leq 2{,}000$).

2.4 Phonotactic modeling of languages [9]

A classical approach of language identification is applied to quantify native-likeness. In the classical approach, a continuous phoneme recognizer of a specific language, e.g. English, was applied to a given utterance of any language. Then, the utterance was represented in a forced way as a sequence of English phonemes. Languages of interest were modeled separately as phoneme N-gram using the forced English phonemes. If we consider a special case of N=1, the model becomes phoneme distribution. After converting the 30-sec long utterance of each participant into its posteriorgram, we can calculate the averaged posterior probability of the n classes ($2 \leq n \leq 2{,}000$), which directly corresponds to distribution of the n classes.

2.5 Elastic Net regression [10]

In this study, for feature selection and for prediction, Elastic Net regression is used for a specific purpose. Mathematically speaking, Elastic Net regression is a combination of Ridge regression [11] and Lasso regression [12], i.e. a combined use of L1 norm and L2 norm as regularization terms for weights. Value normalization is also done for each feature. Because of these, weight coefficients attached to features of less predictability become zero and this is why the function of Elastic Net is said to be prediction based on feature selection.

3 Speech features extracted for prediction

We introduce three types of features for automatic prediction, 1) those derived only from speech acoustics with signal processing techniques, 2) those derived from posteriorgrams of utterances, and 3) those derived from ASR results of the utterances. They are related to quantity of phonation and/or quality of pronunciation. Since the utterances in the corpus [5] are with unignorable noises, two versions of WSJ-KALDI-based English speech recognizers [13] were trained, one with the WSJ corpus only and the other with WSJ and its noisy versions, where three levels of noises (SNR=10, 30, 40 [dB]) were added and all the clean and noisy utterances were used together to train a noise-robust speech recognizer. Results will be shown separately for the baseline recognizer and the noise-robust recognizer.

Even with the noise-robust recognizer, the recognition accuracy was very low and we decided not to use recognized words as they were. However, we extracted some statistics from the recognized words, which were tested for prediction.

3.1 Features derived with signal processing

Following a previous study [14], envelope-based syllable detection was used, which is provided as Praat script [6]. Then, speaking rate was calculated as

$$\text{speaking rate} = \frac{\text{\#syllables}}{\text{total duration of phonation}}$$

The denominator is defined as the utterance length minus its entire duration of pauses. Speaking rate dose not tell anything on how many silent frames are found in the utterance. We introduced a similar but different feature of phonation ratio [15] as

$$\text{phonation ratio} = \frac{\text{total duration of phonation}}{\text{utterance length}}$$

3.2 Features derived from posteriorgrams

From the posteriorgram of each utterance, after pause removal, the following three types of features are calculated automatically.

3.2.1 Average of maximum posterior probabilities [15]

Here, from a given posteriorgram, we detect the maximum posterior probability for each time and it is averaged over time. The higher the average is, the more distinct pronunciation the utterance is made with. The number of phoneme classes n can vary from 2 to 2,000.

3.2.2 Averaged posterior distribution as fine phoneme distribution

As discussed in Section 2.4, the averaged posterior vector can be viewed as the distribution of phonemes. Since the utterance from each participant is so long as 30 sec, a variety enough of phonemes are supposed to exist and the averaged posterior vector can characterize native-likeness of each participant. The number of phoneme classes n can vary from 2 to 2,000, and the average posterior vector is directly used for prediction.

3.2.3 Posterior gap between a participant and native speakers

For each participant, we calculate the averaged posterior vector. Since we have 10 native speakers in the participants, we calculate distance from a participant to each native speaker, 10 gaps in total. The Bhattacharyya distance

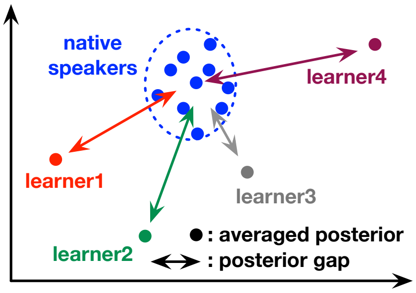

Fig. 2 Averaged posterior and posterior gap

Table 1 Prediction with quantity features

ASR	DNN	1)	2)	3)	4)	corr.
w/o	—	0.768	1.333	—	—	0.819
with	clean	0.744	1.245	0.000	0.182	0.821
with	noise	0.765	1.281	0.000	0.097	0.817

is used again as metric with variable dimension n. These gaps quantify native-likeness of each participant more directly and the averaged gap is used for prediction of fluency. Figure 2 visualizes the averaged posterior and the posterior gap. The former characterizes quality of pronunciation, location in the feature space, and the latter characterizes relative distances to the 10 native speakers.

3.3 Features derived from ASR results

We tested two versions of WSJ-KALDI-based speech recognizers, i.e. clean model and noise-robust model on all the 100 utterances. The results showed that 29.5 % and 32.1 % as correct recognition rates, respectively for the two models. Since these rates are very low, we did not use any features that characterize lexical identity of the recognized results. However, some statistics are supposed to be calculated rather adequately and they are used as feature for fluency prediction.

3.3.1 Correct recognition rate

For this study, the fifth author provided correct transcripts of all the 100 utterances, and with them, we can calculate the correct recognition rate for each participant. Although transcripts of spontaneous utterances are generally unavailable, we tentatively use the correct recognition rate as feature for prediction. The prediction performance with transcripts is just for reference, which may be used as upper limit of prediction.

3.3.2 Total number of words in ASR results

Phonation ratio characterizes how continuously a participant speaks, and that acoustically. It is possible to derive a similar but different feature lexically from recognition results. The recognition performance is surely low, but the total number of words in the recognition results may be effective for prediction.

3.3.3 Size of vocabulary in ASR results

It is easily expected that poor participants may utter the same words repeatedly. This expectation led us to use the number of *different* words, size of vocabulary, for prediction [16].

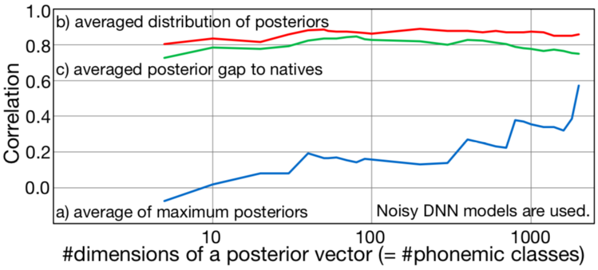

Fig. 3 Correlation as function of posterior dim

Table 2 Prediction with quality features

ASR	DNN	a)	b)	c)	d)	corr.
w/o	clean	0.124	0.365	-0.624	—	0.903
w/o	noise	0.233	0.272	-0.753	—	0.917
with	clean	0.000	0.254	-0.491	0.628	0.922
with	noise	0.045	0.214	-0.549	0.537	0.921

4 Automatic prediction of fluency

4.1 Prediction with quantity features

In [5], features related to smoothness or quantity of phonation were manually extracted. Among the automatically extracted features, we regard 1) speaking rate, 2) phonation ratio, 3) total number of words, and 4) size of vocabulary as features related to quantity of phonation. Table 1 describes results of Elastic Net regression with these features for fluency prediction, where correlations between the averaged fluency scores over the 10 native raters and the machine scores are calculated based on 5-fold cross-validation. In the table, clean and noise mean the two types of ASR models, and the three values assigned to each kind of feature is the weight coefficients calculated for that feature. Clearly shown, phonation ratio and speaking rate are very effective for prediction. The performance is higher than the average of one-to-one inter-rater correlations but much lower than the average of one-to-others correlations.

4.2 Prediction with quality features

The other features, a) average of maximum posteriors, b) averaged distribution of posteriors, and c) posterior gap to natives, are tested with Elastic Net regression. d) correct recognition rate is also tentatively considered. Figure 3 shows correlations as a function of the dimension n of posterior probabilities calculated with noisy DNN model. For a) and c), *feature* correlations are plotted while, for b), *model* correlations (*prediction* correlations) are shown with Elastic Net regression. Correlations with b) and c) are maximized around n=50, while those with a) seem to be higher with larger n, but still lower than those with b) and c). From these results, we select 50 as n and use it for testing all the quality features.

Table 3 Prediction with all the features

ASR	DNN	c)	1)	a)	2)	d)	corr.
w/o	clean	-0.589	0.224	0.112	0.364	—	0.906
w/o	noise	-0.748	0.264	0.255	0.231	—	**0.925**
with	clean	-0.580	0.194	0.131	0.334	—	0.906
with	noise	-0.715	0.242	0.233	0.200	—	0.923
with	clean	-0.476	0.192	0.000	0.311	0.602	0.923
with	noise	-0.543	0.276	0.033	0.239	0.561	0.928

Table 2 describes results of Elastic Net regression with the quality features for fluency prediction. As b) is a multivariate feature, its weight means the largest weight among the n dimensions. Clearly shown, c) and b) are very effective for prediction. It is very surprising to us that the correlation with the quality features only even without ASR overcomes the average of one-to-others correlations (0.873), and is comparable to the maximum (0.910). This claims that the trained model is comparable to the most stable and reliable human rater.

4.3 Prediction with all the features

Table 3 describes results of Elastic Net regression with all the features. Only the top four features in the case of noisy DNN but without ASR are shown, also in other cases with or without d) correct recognition rates. In the table, the top four features are c) averaged posterior gap to natives, 1) speaking rate, a) average of maximum posteriors, and 2) phonation ratio. i.e. two quality features and two quantity features. In the table, very high usability of the quality features is shown again and, even without ASR, the trained model gives a higher correlation of 0.925 than the maximum of one-to-others correlations (0.910).

4.4 Discussion and future directions

In this paper, we tried to predict subjective scores of fluency. What we found is that the fluency scores can be much more highly predicted with quality features than with quantity features. This result implies that 1) judgments of the 10 native raters were rather biased to the quality of pronunciation, which is logically independent of smoothness and fluidity in utterances, or 2) quantity features and quality features are highly correlated and the latter were extracted with higher accuracy. We're interested in another kind of fluency scores, given by expert raters. With expert rating, we may obtain some different results.

5 Conclusions

This paper presented research results of a joint project between UTokyo and UCL, where automated scoring of fluency was investigated. Since the L2 corpus prepared for development was not large, we tested classical machine learning techniques with recently proposed speech representations such as posteriorgram with variable granularity. Experiments showed a correlation of 0.925 to the perceived fluency, which was higher than the maximum inter-rater (one-to-others) correlation (0.910).

Reference

[1] M. Eskenazi, “An overview of spoken language technology for education,” *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] T. Kawahara and N. Minematsu, “Computer-Assisted Language Learning (CALL) based on speech technologies,” *IEICE Trans. Info. Sys.*, vol. J96-D, no. 7, pp. 1549–1565, 2013.

[3] L. Chen, L. Davis, K. Zechner, C. M. Lee, S.-Y. Yoon, M. Ma, K. Evenini, R. Mundkowsky, X. Wang, C. Lu, A. Loukina, C. W. Leong, J. Tao, and B. Gyawali, “Automated scoring of nonnative speech using the SpeechRater v.5.0 engine,” *ETS Research Report Series*, vol. RR-18, no. 10, pp. 1–31, 2018.

[4] T. Isaacs, “Fully automated speaking assessments: changes to proficiency testing and the role of pronunciation,” in *The Routledge handbook of contemporary English pronunciation*, O. Kang, R. I. Thomson, and J. Murphy, Eds. Routledge, 2018, pp. 570–584.

[5] K. Saito, M. Ilkan, V. Magne, M. N. Tran, and S. Suzuki, “Acoustic characteristics and learner profiles of low-, mid-and high-level second language fluency,” *Applied psycholinguistics*, vol. 39, no. 3, pp. 593–617, 2018.

[6] P. Boersma and D. Weenink, Praat:doing phonetics by computer (Version 6.1.03)[computer software], 2019. [Online]. Available: http://www.praat.org

[7] Y. Kashiwagi, C. Zhang, D. Saito, and N. Minematsu, “Divergence estimation based on deep neural networks and its use for language identification,” in *Proc. ICASSP*, 2016, pp. 5435–5439.

[8] J. H. Ward,“Hierarchical grouping to optimize an objective function,” *Journal of the American Statistical Association*, vol. 58, pp. 236–244, 1963.

[9] P. Mateika, P. Schwarz, J.Cernocky, and P. Chytil, “Phonotactic language identification using high quality phoneme recognition,” in *Proc. INTERSPEECH*, 2005, pp. 2237–2240.

[10] H. Zou and T. Hastie, “Regularization and variable selection via the Elastic Net,” *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[11] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Journal of Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[12] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., “The Kaldi speech recognition toolkit,” in *Proc. ASRU*, 2011.

[14] L. Fontan, M. L. Coz, and S. Detey, “Automatically measuring L2 speech fluency without the need of ASR: a proof-of-concept study with Japanese learners of French,” in *Proc. INTERSPEECH*, 2018, pp. 2544–2548.

[15] A. Yasukagawa, S. Ando, E, Konno, Z. Lin, Y. Inoue, D. Saito, N. Minematsu, and K. Saito, “An experimental study of automatic scoring of fluency of spontaneous English utterances by Japanese learners,” *Proc. Spring Meeting of Acoustical Society of Japan*, 2020.

[16] H. Hilton, “The link between vocabulary knowledge and spoken L2 fluency,” *Language Learning Journal*, vol. 36, no. 2, pp. 153– 166, 2008.