Backend Design

---- Group 5: Yoyo D., Jiatong S., Nan H., Gary X., Rafa de la TO

Currently, we can split backend development into three stages, including algorithmic refinement, evaluation metrics, front-back interaction and stress the model.

Algorithmic Refinement

In this part, we added the Prosodic Model in our backend system.

- **Duration Model**

We use BiLSTM for phonetic duration modeling. The input of the model is a one-hot phonetic sequence and the target is to predict a sequence of duration. Comparing to the model defined in, the framework can extract context information on phonemes and adapt to unmet words with the knowledge from the training process.

- **Prosodic Scoring**

Firstly, the speech is processed to the forced alignment according to the reference text. Then, the aligned speech is split into word-level. Meanwhile, the transcript split to word-level is fed into the duration model to compute a template duration relationship for the speech. With the normalization technique introduced, the word-level alignments are rescaled.

The scoring method on word-level is shown below:

$$Prosody = -(\sum_{p \in P} \delta(p)) \,/\, |P|$$

$$\delta(p) = Max(abs(|(D_{True} - D_{Pred})|) - Mean(E_p) - Std(E_p), 0)$$

where δ(p) is the phone-level scoring function that measures the error exceeding the threshold. The $D_{True} - D_{Pred}$ represents the difference between the true label and the prediction label. The E_p is the absolute prediction erroneous set observed from the duration model training process.

- **Prosodic GOP (PGOP)**

Because the speech comprehension process is partially based on the duration prosody, it can be assumed that the pronunciation with longer duration plays a more important role in determinization of the integrity (if the word can be properly recognized) and naturalness (if the word is spoken fluently) for the language.

Since the duration model can output a reference duration length for a word-level sequence, a Prosodic GOP is proposed as follows:

$$PGOP(p) = \alpha_p \cdot sGOP(p) + \beta_p \cdot \delta(p) * sGOP(p) + c_p$$

where α_p, β_p, and c_p are estimated with linear regression. The construction of the PGOP based on an assumption that a wrong pronunciation with the wrong duration does more harm than a wrong pronunciation with a better duration. Therefore, the PGOP is a context-related scoring method instead of an independent pronunciation error detector.

- **Evaluation:**

Table The PGOP Scoring Experiments

Method / Rater	Pearson-1	Pearson-2	Spearman-1	Spearman-2	MIC-1	MIC-2
Rater 1	/	0.573	/	0.573	/	0.276
GOP	0.425	0.297	0.370	0.315	0.182	0.143
SGOP	0.452	0.287	0.409	0.304	0.212	0.173
PGOP	**0.511**	**0.347**	**0.420**	**0.365**	**0.232**	**0.187**

the parameters for PGOP is estimated with 70% of the CALL_2K. The α_p, β_p, and c_p are 0.3883, -0.0049 and 51.3421. The result is shown in this Table. On all the measures, the PGOP outperforms the GOP method and the SGOP method. It validates the prosodic importance to the human's language comprehension process.

Features added

We will add some features in the feedback part including detecting if the user reads the wrong stress on vowels. And we implement the detection of wrong phones which will be shown to the users in feedback.

Front-back interaction

We will send the feedback in the form of JSON to the frontend and display it to users, which contains: the length and pronunciation score of each phone, and the information of wrong pronunciation wrong stress and the score of fluency.

Stress the model

When we try to randomly speak to our system, the system will still give some points like 5/100. Sometimes it will up to 10~20 points which is crazy. And when we input the audio we download from the e-dictionary, the score is around 85~90, which should be 100 points. And also the time consumed for each try is slow and we will improve it later.

If our system fails

- Severe consequences: the wrong feedback to the user may mislead our users and thus ruin the original purpose of our system.
- Mild consequences: the perfect audio can't get a full score which is okay since those with perfect pronunciation should already have enough confidence in their Language skills.
- No consequences: When our system is not that stable and encountered with a sudden shutdown, users may just restart it, which will cause nothing except some madness. ^-^