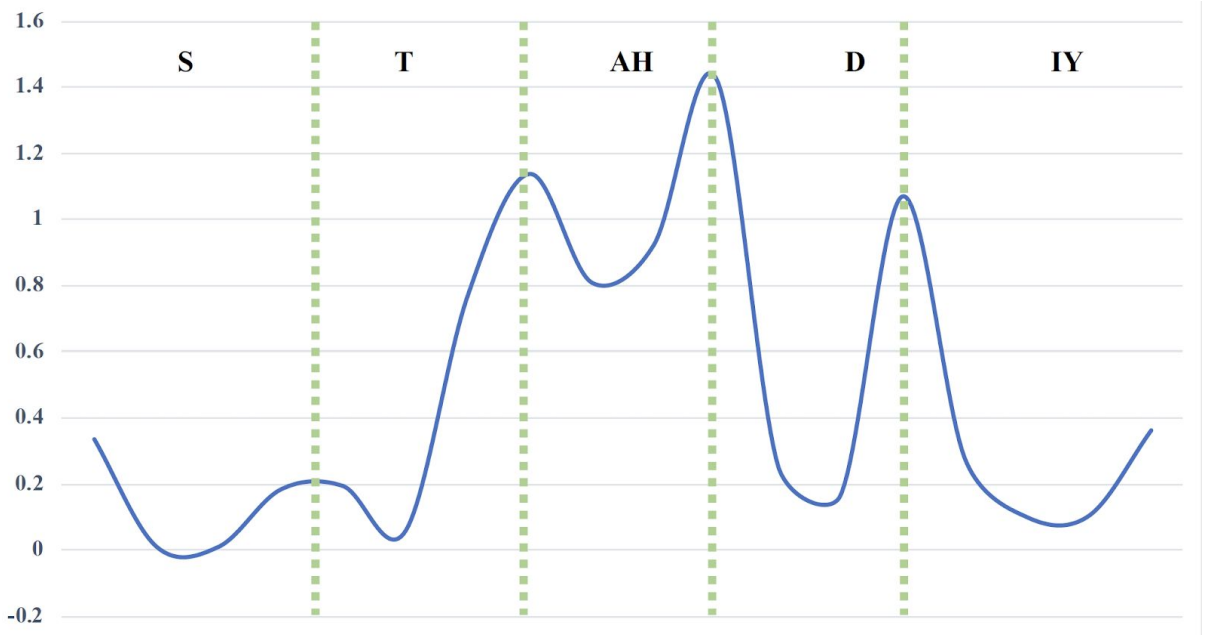
Further Implementation

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

This time, we further refine the backend with transition aware GOP and self-attention duration model.

For the previous sGOP, we used a fixed segment of the specific phone so as to skip the speech transition effect when speaking through phonemes. However, we found some bad cases that the phonetic transition may not occur just at the sides of the phonetic segment. It might be “skewed”. So here, we propose the transition aware GOP using the entropy of post-probability to identify the salient part of the phonetic segment. For each frame, a lower entropy of post-probability states that the acoustic model is more confident in the prediction. The fluctuation in post-probability entropy is likely to be the transition between phonemes. The Figure is an example of the post-probability entropy on the word "Study" using our acoustic model.



To compute the Transition Aware GOP, we first reformulate the GOP to frame-wise GOP using conditional independence as Formula, where the GOP(a, t) is the GOP on t frame for phoneme a.

$$\mathrm{GOP}(a,t) = log(\frac{p(o_t|a)\cdot p(a)}{\sum_{a'\in A} p(o_t|a')p(a')})$$

Then, we compute the frame-wise post-probability entropy defined as Formula \ref{eq:entropy}, where p(a',t) is the post-probability of the phone a' at time t. Then, we weight the frame-wise GOP with the reciprocal of entropy as shown in Formula

$$\text{TransitionAware}(a) = \sum_{t}^{N} \frac{\frac{1}{-\sum_{a' \in A} p_{a',t} \cdot log(p_{a',t})}}{\sum_{t}^{N} \frac{1}{-\sum_{a' \in A} p_{a',t} \cdot log(p_{a',t})}} \cdot \text{GOP}(a,t)$$

For this transition aware model, we observe a 0.03 correlation improvement then sGOP.
Phonetic Duration is an essential factor in speech articulation and also contributes to speech comprehension. It is shown to benefit many speech processing tasks (e.g. Text-to-speech, ASR). Since it is a sequential problem, previous literature explored to use statistical graphic model (e.g. HMM). Recently, Neural networks have also been introduced to the problem and raised reasonable improvement comparing to conventional methods. In this section, we propose to use the multi-head self-attention mechanism to model the phonetic duration.

We firstly pass the reference text to the text-to-phone converter and use phone-id sequences as our model input. Then we add the speed and positional information. The speed is referred to the average duration of phones in the given speech. We first remove the un-voicing part using voice activity detection (VAD) and then divide the remaining time by the number of phones from the reference text. The self-attention encoder follows the definition of as shown in Formula but with a local diagonal Gaussian matrix to favor local information.

We use DNN and LSTM as baselines. Since the DNN cannot model the sequential information, we adopt a sliding window of seven to input the sequential information to the DNN. For the evaluation, we use the TIMIT dataset, which is a dataset that has phonetic duration annotation. The result shows in the following table by mean absolute error (MAE), indicating the power of our proposed model:

Model	MAE
DNN	89.86ms
LSTM	42.60ms
Self-attention	32.4ms