ASR-Free Pronunciation Assessment

Sitong Cheng
Zhixin Liu

Introduction

Goodness of Pronunciation (GOP)

- GOP is based on the posterior probability on the correct phone, given the speech segment of that phone

$$GOP = \frac{1}{M}\sum_{i}^{M} \ln p(q_i|o_i),$$

q_i = i-th phone in the speech segment
o_i = corresponding speech segment
M = total number of phones in the speech segment

- If given a phone sequence q and the corresponding speech signal o, and assume the alignment is perfect

$$p(\mathbf{o}|\mathbf{q}) = \frac{1}{M}\sum_{i} \ln p(o_i|q_i)$$

$$p(\mathbf{o}|\mathbf{q}) = \frac{1}{M}\sum_{i=1}^{M} \ln \frac{p(q_i|o_i)p(o_i)}{p(q_i)}$$

ignored

However ……

Problems

- There is no guarantee that a worse pronunciation will achieve a smaller posterior!

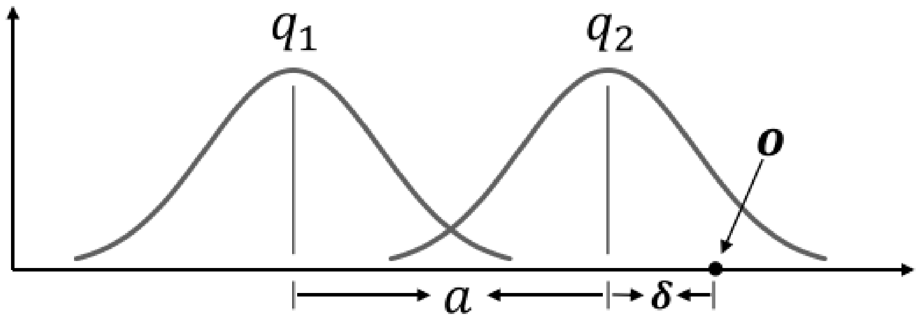
- perfect pronunciation of q_2 :

$$p(q_2|o) = \frac{1}{1+e^{-a^2}}$$

- non-native speaker pronounce q_2 at a position o

$$p(q_2|\mathbf{o}) = \frac{e^{-\delta^2}}{e^{-\delta^2} + e^{-(a+\delta)^2}} = \frac{1}{1 + e^{-(a^2+2a\delta)}}$$

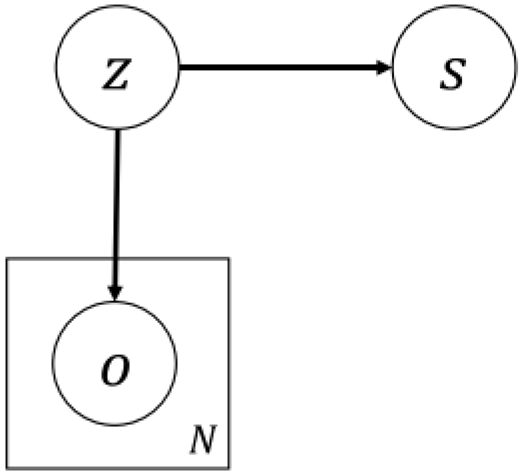
If δ > 0, the posterior essentially increases. This means that a non-native speaker obtains a better GOP than a native speaker.



Solutions:

ASR-free scoring

- We conjecture that it is the marginal part $p(o)$ that solves the phone competition. Since $p(o)$ concerns neither phones nor words, it is called an ASR-free scoring.



- Method

  We cannot directly use $p(o)$, since $p(o)$ is quite noisy

  $p(o) \rightarrow p(z) \ \rightarrow p(s)$

Three Marginal models

$p(o) \rightarrow p(z)$

i-vector model

Normalization flow

All vectors are trained with je-1520

Discriminative NF

Prediction model

$p(z) \rightarrow p(s)$

- With SVR, we predict the score s directly, which can be regarded as a special form of the prediction distribution $p(s|z)$ where all the probability mass concentrates in a single value.
- SVR train with je-1520, test with je-380.

Information fusion

Score fusion

$$s^* = \lambda p(\mathbf{q}|\mathbf{o}) + (1-\lambda) \arg\max_s \{p(s|\mathbf{o})\}$$

$$s^* = \lambda p(\mathbf{q}|\mathbf{o}) + (1-\lambda)\gamma(\mathbf{o})$$

$\gamma(\cdot)$ is the prediction function implemented by SVR

Feature fusion

- we treat the GOP score $p(q|o)$ as a feature and combine it with the latent representation z, and then build the SVR model.

Experiments & Results

- All results use PCC

Basic results

	Human	GOP	GMM	NF
PCC	0.550	**0.614**	-0.065	-0.131

ASR-free scoring

	i-vector + SVR	NF + SVR	DNF + SVR
PCC	0.434	0.441	**0.462**

Information fusion

	Score-fusion	Feature-fusion
GOP + i-vector	0.640 ($\lambda = 0.38$)	0.625
GOP + NF	0.663 ($\lambda = 0.34$)	0.656
GOP + DNF	0.676 ($\lambda = 0.36$)	0.667

Conclusion

- Our theoretical study shows that this scoring approach offers an interesting correction for the phone-competition problem of GOP, and empirical study demonstrated that combining the GOP and this ASR-free approach can achieve better performance than the GOP baseline.

Thank you !