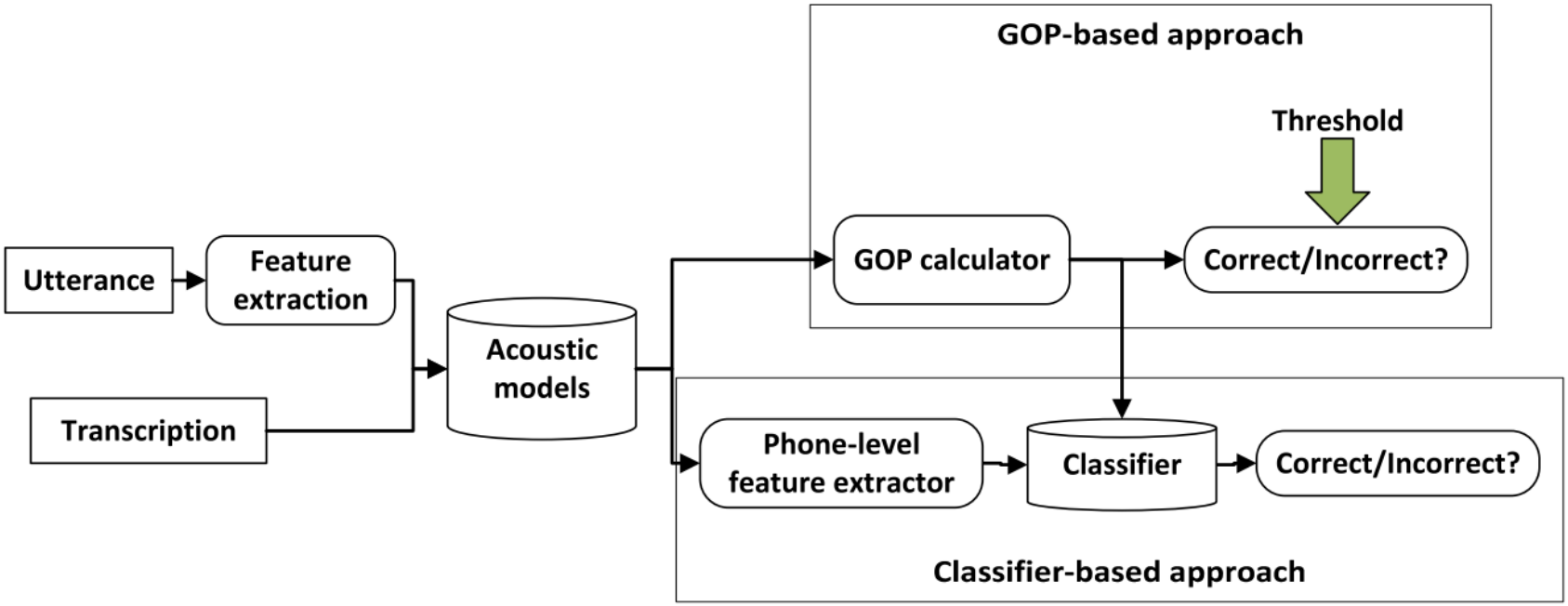
发音良好度 (GOP) 基于 Kaldi 的实现

张俊博

小米科技园，北京
2019-10-26

A Typical Mispronunciation Detection System



(Hu et al., 2015)

GMM-based Goodness Of Pronunciation (GOP)

In the conventional GMM-HMM based system, GOP was first proposed in (Witt et al., 2000).
Defined as the duration normalised log of the posterior:

$$GOP(p) = \frac{1}{t_e - t_s + 1} \log p(p|\mathbf{o}) \tag{1}$$

- $\mathbf{o}$ the input observations
- p the canonical phone
- t_s, t_e the start and end frame indexes

GOP-GMM Calculation

$$\log p(p|\mathbf{o}) = \frac{p(\mathbf{o}|p)p(p)}{\sum_{q\in Q} p(\mathbf{o}|q)p(q)} \approx \frac{p(\mathbf{o}|p)}{\sum_{q\in Q} p(\mathbf{o}|q)} \tag{2}$$

Q the whole phone set

How To Calculate

Numerator Forced alignment

Denominator Viterbi decoding with an unconstrained phone loop, assuming $\sum_{q_t\in Q} p(\mathbf{o}|q_t) \approx \max_{q_t\in Q} p(\mathbf{o}|q_t)$

An implemented example (deprecated):
`https://github.com/jimbozhang/kaldi-gop`

DNN-based GOP

In the NN-HMM based system, GOP was defined as the log phone posterior ratio between the canonical phone and the one with the highest score (Hu et al., 2015).

$$LPP(p) = \log p(p|\mathbf{o}; t_s, t_e) \tag{3}$$

$$GOP(p) = \log \frac{LPP(p)}{\max_{q \in Q} LPP(q)} \tag{4}$$

GOP-NN Calculation

$$LPP(p) \approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log p(p|o_t) \tag{5}$$

$$p(p|o_t) = \sum_{s \in p} p(s|o_t) \tag{6}$$

s the senone label

$\{s|s \in p\}$ the states belonging to those triphones whose current phone is p

Phone-level Feature

Extract a feature vector for each phone segment $\mathbf{o}_i$ for supervised training. The phone-level feature is defined as:

$$[LPP(p_1), \cdots, LPP(p_M), LPR(p_1|pi), \cdots, LPR(p_j|p_i), \cdots]^T \tag{7}$$

where the Log Posterior Ratio (LPR) between phone p_j and p_i is defined as:

$$LPR(p_j|p_i) = \log p(p_j|\mathbf{o}; t_s, t_e) - \log p(p_i|\mathbf{o}; t_s, t_e) \tag{8}$$

An implemented example (**临时地址**):
`https://github.com/jimbozhang/kaldi/tree/gop/egs/gop`

Reference

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.

Wenping Hu, Yao Qian, Frank K. Soong, and Yong Wang,
"Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers,"
Speech Communication, vol. 67, no. January, pp. 154–166, 2015.