Automatic Scoring of L2 English Speech Based on DNN Acoustic Models with Lattice-Free MMI

Dean Luo(✉), Mingxiang Guan, and Linzhong Xia

Shenzhen Institute of Information Technology, Shenzhen, China
luoda@sziit.edu.cn

Abstract. This paper proposed improved automatic scoring methods for L2 English speaking tests based on acoustic models with lattice-free Maximum Mutual Information (MMI). Deep Neural Network (DNN) acoustic modeling with lattice-free MMI is the state-of-the-art technology in speech recognition because of its effectiveness in sequential discriminative training. Novel Goodness of Pronunciation (GOP) implementations based on lattice free MMI were proposed to improve the performance of automatic scoring for L2 English speech tests. Sequential acoustic weights during forced-alignment and posteriors based on Forward-Backward Algorithm with lattice free MMI acoustic models were used to improved GOP based automatic scoring. Experimental results show that our proposed lattice free MMI based methods outperform conventional regular DNN based automatic scoring methods.

Keywords: Automatic scoring · L2 speech evaluation · Goodness of pronunciation · Lattice free MMI · DNN acoustic models

1 Introduction

English speaking tests have been incorporated in high-stake tests such as National College English Test (CET), College Entrance Examination and senior high school entrance examination. It is very time-consuming and labor intense to evaluate L2 speech manually in large-scale tests such as CET. Human raters are usually required to have some expertise in both phonetics and language education. Therefore, it is not practical to assess L2 speech manually for large-scale tests.

Computer-Aided Language Learning (CALL) based on automatic speech recognition (ASR) has been very active for the past two decades. One application of CALL is automatic scoring based on ASR which uses machine learning algorithms to learn human experts' scoring strategies from features extracted with speech processing [1–5].

Dramatic improvements have been reported with ASR based on Deep Neural Networks (DNN) in recent years [6, 7]. Since 2011, we have been working on improvements of automatic scoring methods for English speaking tests [8–10]. In this paper, we proposed two novel implementations of Goodness of Pronunciation scores to better utilize criminative training power of lattice free MMI based DNN acoustic

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2021
Published by Springer Nature Switzerland AG 2021. All Rights Reserved
M. Guan and Z. Na (Eds.): MLICOM 2020, LNICST 342, pp. 113–122, 2021.
https://doi.org/10.1007/978-3-030-66785-6_13

models. Experimental results show that our proposed methods outperform conventional DNN approaches for automatic scoring in high-state English speaking tests.

2 Automatic Scoring for L2 Speech

2.1 Goodness of Pronunciation

The Goodness of Pronunciation (GOP), which is defined as phone level confident sore extracted for ASR, is often used for evaluate learners' pronunciation and often used for automatic scoring [11]. The GOP score can be defined as the following,

$$\begin{aligned} \mathrm{GOP}(p) &= \log(\mathrm{p}(p|\mathbf{o})) \\ &\approx \log \frac{\mathrm{p}(\boldsymbol{o}|p)\mathrm{p}(p)}{\max\{q \in Q\} p(\boldsymbol{o}|q)p(q)} \\ &\approx \log \frac{\mathrm{p}(\boldsymbol{o}|p)}{\max\{q \in Q\} \mathrm{p}(\boldsymbol{o}|q)} \end{aligned} \tag{1}$$

where the probability $\mathrm{p}(p|\mathbf{o})$ is the posterior of phoneme p given speech segment feature **o**, Q represents the whole set of all phonemes. The numerator of Eq. (1) is a likelihood that can be attained with phone-level GMM- HMM forced alignment, and the denominator is the maximum likelihood of any phonemes recognized by HMM with a grammar model generated through a phone-loop grammar network.

2.2 DNN Based GOP Implementation

The most popular DNN based ASR engines usually train a neural network to output HMM state-level probabilities [12]. For state likelihoods, the outputs of DNN are divided by the priors attained through acoustic model training.

For DNN based acoustic models, GOP can be calculated using the average state posteriors, which is usually the softmap output of DNN models [7]:

$$\mathrm{GOP}(p) = \mathrm{p}(\mathrm{p}|t_s, t_e; \boldsymbol{O}) = \frac{1}{t_e - t_s} \sum_{t_s}^{t_e} \mathrm{p}(s_t|\boldsymbol{o}_t), \tag{2}$$

where $p(s_t|\boldsymbol{o}_t)$ is the state-level posterior that is attained from the output of DNN, $\boldsymbol{o}_t$ is the segment of acoustic feature at time t, t_s is the start and t_e are the end time of the acoustic feature of phoneme p, which can be obtained with forced alignment. We refer this baseline definition of DNN based GOP as GOP1.

As mentioned above, with DNN-HMM hybrid acoustic models, the state-level posteriors are converted to state-level quasi-likelihoods by dividing with the priors of the HMM states. Therefore, the numerator and denominator of Eq. (1) can be calculated with forced alignment:

$$\mathrm{P}(\boldsymbol{o}|p) \approx \frac{1}{t_e - t_s} \sum\nolimits_{t_s}^{t_e} \mathrm{p}(s_t|\boldsymbol{o}_t)/\mathrm{p}(s_t), \tag{4}$$

$$p(\boldsymbol{o}|q) \approx \frac{1}{t_e - t_s} \sum_{t_s}^{t_e} \max\left\{s_t^* \in S\right\}(P(s_t^*|\boldsymbol{o}_t)/p\left(s_t^*\right)), \tag{5}$$

where S represents the full set of all HMM states also called "senones", $\mathrm{p}(s_t)$ the state-level prior of the state at time t with is attained during DNN-HMM acoustic model training. If we apply Eqs. (4) and (5) to Eq. (1), we get another GOP score, and we call it GOP2 hear after. GOP2 is more robust than GOP1, since noises due to acoustic mismatches appear in the numerator and denominator of Eq. (1) . GOP1 and GOP2 are used as baseline automatic scores in this study.

3 Automatic Scoring Based on Lattice Free MMI

3.1 Lattice Free MMI Acoustic Modeling

The objective function for estimating HMM parameters of GMM-HMM or DNN-HMM in speech recognition is define as:

$$F_{ML} = \sum_{r=1}^{R} \log P_\theta(O_r|W_r)P(W_r) \tag{6}$$

$$= \sum_{r=1}^{R} \log \sum_{s \in w_r} \prod_{t=0}^{T_r - 1} P(s_{t+1}|s_t)P(O_r(t)|s_t) \tag{7}$$

where θ is the set of HMM parameters, R is the number of all training utterances, O_r is the r^{th} utterance with length T_r, and W_r is all the possible sequences given the transcription.

Maximum Mutual Information (MMI) is can be considered as a discriminative objective function which is used to maximize the probability of the reference transcriptions, while minimizing the probability of all other alternatives:

$$F_{MMI} = \sum_{r=1}^{R} \log \frac{P_\theta(O_r|W_r)P(W_r)}{\sum_{\widehat{W}} P_\theta\left(O_r|\widehat{W}\right)P\left(\widehat{W}\right)} \tag{8}$$

$$= \sum_{r=1}^{R} \left[\log P_\theta(O_r|W_r) + logP(W_r) - \log \sum\nolimits_{\widehat{W}} P_\theta\left(O_r|\widehat{W}\right)P\left(\widehat{W}\right)\right] \tag{9}$$

If we take the gradient with respect to parameter θ:

$$\nabla_\theta F_{MMI}[\theta] = \nabla_\theta \sum_{r=1}^{R} \left[\log P_\theta(O_r|W_r) + logP(W_r) - \log \sum\nolimits_{\widehat{W}} P_\theta\left(O_r|\widehat{W}\right) P\left(\widehat{W}\right) \right] \tag{10}$$

Since $logP(W_r)$ is independent of θ, we use $\nabla_\theta logP(W_r) = 0$,

$$\nabla_\theta F_{MMI}[\theta] = \sum_{r=1}^{R} \left[\nabla_\theta \log P_\theta(O_r|W_r) - \nabla_\theta \log \sum\nolimits_{\widehat{W}} P_\theta\left(O_r|\widehat{W}\right) P\left(\widehat{W}\right) \right] \tag{11}$$

$$= \sum_{r=1}^{R} \left[\nabla_\theta \log P_\theta(O_r|W_r) - \frac{\sum_{\widehat{W}} P\left(\widehat{W}\right) P_\theta\left(O_r|\widehat{W}\right) \nabla_\theta \log P_\theta\left(O_r|\widehat{W}\right)}{\sum_{\widehat{W}} P_\theta\left(O_r|\widehat{W}\right) P\left(\widehat{W}\right)} \right] \tag{12}$$

If we define state occupancy probability (word sequence conditioned state posterior) as:

$$\gamma_{rt}(s|W) = \frac{\sum_{s' \in s} P_\theta(O_r, s'|W)}{P_\theta(O_r|W)} = P_{\theta,t}(s|O_r, W) \tag{13}$$

where s is the state sequence and W is word sequence, our gradient equation becomes,

$$\begin{aligned} \nabla_\theta F_{MMI}[\theta] = \sum_{r=1}^{R} \sum_{t=1}^{T} \sum_{s} \nabla_\theta \log P_\theta(O_t|s_t) \Big(& \gamma_{rt}(s|W_r) \\ & - \frac{\sum_{\widehat{W}} P\left(\widehat{W}\right) P_\theta\left(O_r|\widehat{W}\right) \gamma_{rt}\left(s|\widehat{W}\right)}{\sum_{\widehat{W}} P_\theta\left(O_r|\widehat{W}\right) P\left(\widehat{W}\right)} \Big) \end{aligned} \tag{14}$$

$$= \sum_{r=1}^{R} \sum_{t=1}^{T} \sum_{s} \nabla_\theta \log P_\theta(O_t|s_t) (\gamma_{rt}(s|W_r) - \gamma_{rt}(s)) \tag{15}$$

where $\gamma_{rt}(s)$ is the general state posterior and computed by Forward-Backward algorithm with denominator graph of the lattice free MMI models, $\log P_\theta(O_t|s_t)$ can be obtain through neural network output, and $\gamma_{rt}(s|W_r)$ can be calculate with numerator graph of LF-MMI models.

Povey et al. [13] used MMI training for lattice free MMI models using a full denominator graph using a phone-level language model instead of a word-level language model for the denominator graph. The conducted the denominator computation on GPU instead of CPU. The phone-level language model for the denominator graph was a pruned n-gram language model trained with the phone-level alignments of the training transcription. Also, the numerator graph doesn't use the composite HMM and instead uses a special acyclic which could better utilize the alignment information from previous acoustic models. The numerator graph of the regular Lattice-free MMI method can be considered as an expanded version of the composite HMM. The amount

of the self-loops expansion for each utterance is determined by the alignment (i.e. there are no self-loops).

3.2 Forward-Backward Algorithm

Forward-Backward (FB) algorithm is widely used to calculate state occupancy probabilities for HMM parameter estimation. $\gamma_{rt}(s|W_r)$ and are calculated through Forward-Backward algorithm with numerator and denominator graphs.

The forward probability of FB algorithm is the joint probability of observing first t speech vectors and being in state j at time t, given some model M:

$$\alpha_j(t) = P(o_1, \ldots, o_t, s_t = j|M). \tag{16}$$

It can be efficiently calculated by the following recursion

$$\alpha_j(t) = \left[\sum_{i=2}^{N-1} \alpha_i(t-1)a_{ij}\right] b_j\big(b_j(o_t)\big) \tag{17}$$

where a_{ij} is transition probability from state i to j, $b_j(o_t)$ is the probability density of speech vector o_t being generated from state j, and N is the total number of states.

The initial conditions for the above recursion are

$$\alpha_1(1) = 1 \tag{18}$$

$$\alpha_j(1) = a_{1j}b_j(o_1) \tag{19}$$

for 1 < j < N and the final condition is given by

$$\alpha_N(T) = \sum_{i=2}^{N-1} \alpha_i(T)a_{iN}. \tag{20}$$

From the definition of forward probability, $\alpha_N(T)$ is the total likelihood P(O|M), i.e.,

$$\mathrm{P(O|M)} = \alpha_N(T) \tag{21}$$

The backward probability is defined as

$$\beta_j(t) = P(o_{t+1}, \ldots, o_T|s_t = j, M), \tag{22}$$

and can be computed with the following recursion

$$\beta_j(t) = \sum_{i=2}^{N-1} a_{ij}b_j(o_{t+1})\beta_j(t+1) \tag{23}$$

The initial condition is given by

$$\beta_j(T) = a_{iN} \tag{24}$$

and for 1 < j < N and the final condition is given by

$$\beta_1(1) = \sum_{i=2}^{N-1} a_{1j} b_j(o_1) \beta_j(1). \tag{25}$$

By the definitions of the forward and backward probabilities, if we take the product of the two,

$$\alpha_j(t)\beta_j(t) = P(O, s_t = j|M) \tag{26}$$

The state occupation probability in Eq. (13) becomes,

$$P_{\theta,t}(s|O_r, W) = P(s_t = j|O_r, M) \tag{27}$$

$$= \frac{P(O, s_t = j|M)}{\mathrm{P(O|M)}} \tag{28}$$

Substituting Eq. (26) and (21) into Eq. (28) gives:

$$P_{\theta,t}(s|O_r, W) = \frac{\alpha_j(t)\beta_j(t)}{\alpha_N(T)} \tag{29}$$

This demonstrates that state occupation probabilities used for lattice free MMI training can be computed through Forward-Backward algorithm with numerator and denominator graphs.

3.3 GOP Scores Based on Lattice Free MMI

As described in Sects. 3.1 and 3.2, the LF MMI training of DNN-HMM models is compute word conditioned probability which is defined as state occupancy probability given the word sequences using numerator graph, and general state prior calculated though denominator graph with all possible sequences in training data.

To fully utilize the discriminative power of LF-MMI, we implemented two novel GOP scores that can corporate sequential information through LF-MMI. First use word conditioned probability $P_{\theta,t}(s|O_r, W)$ at frame level through forced alignment to substitute DNN out puts $p(s_t|\boldsymbol{o}_t)$ as used in Eq. (2). Capered with the no-linearity characteristics of DNN outputs, conditioned probability $P_{\theta,t}(s|O_r, W)$ incorporates transition probabilities a_{ij} and sequential information W pertaining to transcription which is very suitable for pronunciation assessment of reading-aloud. Therefore, our first implementation of LF-MMI GOP score is given by

$$\text{GOP\_weight}(p) = \text{p}(\text{p}|t_s, t_e; \boldsymbol{O}, \boldsymbol{W}) = \frac{1}{t_e - t_s} \sum_{t_s}^{t_e} P_{\theta,t}(s|O_r, W) \tag{30}$$

Where W can be viewed as the word sequence from transcription. We call this implementation GOP weight, because the word sequenced conditioned probability $P_{\theta,t}(s|O_r, W)$ can be interpreted as state posterior weights during forced alignment.

During the LF-MMI training, the gradient of the denominator of objective function is given by general state posterior $\gamma_{rt}(s)$ which is not dependent of the transcription of a specific utterance. However, $\gamma_{rt}(s)$ is computed with denominator graphs that incorporates all possible sequences in training data. With the success of LF-MMI models in speech recognition, we considered that $\gamma_{rt}(s)$ computed with denominator graph of a LF-MMI model can effectively general sequential information and thus $\gamma_{rt}(s)$ is a reliable state posterior to evaluate pronunciation. As the same with GOP-weight, we substitute $p(s_t|\boldsymbol{o}_t)$ used in Eq. (2) with $\gamma_{rt}(s)$. We call this implementation of GOP as GOP_FB, as it is computed through Forward-Backward algorithm with denominator graph of LF-MMI acoustic models.

$$\text{GOP\_FB}(p) = \text{p}(\text{p}|t_s, t_e; \boldsymbol{O}, \boldsymbol{W}) = \frac{1}{t_e - t_s} \sum_{t_s}^{t_e} \gamma_{rt}(s) \tag{31}$$

As $\gamma_{rt}(s)$ is independent of specific transcripts of a given utterance, GOP-FB can be used for general pronunciation evaluation purposes, not only for reading-aloud, but also other more open task such as retelling or spontaneous conversations.

4 Experiments

4.1 Speech Data and Reference Scores

For our evaluation experiments, we used the L2 speech corpus of Shenzhen High Schools English Speaking Test. We only used the reading-aloud part of the test, in with students are presented with a one-minute long video and required to read out the subtitle of the video. The recordings of this reading-aloud speech used for our automatic scoring experiments.

All together we used 600 senteces uttered by 600 students, including 300 male and 300 female students with various levels of overall proficiency (beginners, intermedium learners and advanced learners). There are 200 learners in each proficiency group.

Three experts were recruited to manually evaluate students' speech and give an overall proficiency score for each utterance. The assessment standard is shown in Table 1.

Table 1. Assessment standard

Score	Scoring standards
5	Fluent and native-like in pronunciation and intonation without any mistakes
4	Fluent and intelligible with minor unnaturalness in pronunciation or intonation. Very few linguistic or phonetic mistakes
3	Have some errors in pronunciation or unnaturalness in intonation, but most part of the speech is intelligible
2	Large amount of pronunciation errors and unnatural intonation, but parts of the speech is still intelligible
1	Severe errors in pronunciation and most part of the speech is unintelligible
0	Completely unintelligible, silence or speaking something unrelated to presented subtitle text

4.2 Acoustic Models

Both conventional DNN-HMM and lattice free MMI models were used for computing different GOP scores. The acoustic models are trained using the Kaldi toolkit [14]. The DNN based models were trained on Librispeech corpus [15].

The acoustic features we used for training monophone and triphone models is mel-frequency cepstral coefficients or MFCCs and their delta and double deltas. The 40-dimensional features are then transformed with Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT). For DNN training, we used time-delay feedforward neural networks with 6 layers. The p of p-norm activations is set to be 2 and the dimensions of the input and output are set to be 2000 and 250 respectively. The initial learning rate is set to 0.005, which was the reduced exponentially to a tenth of the original rate. The number of epochs is set to be 8. The lattice free MMI setups are the same as in [13].

4.3 Experimental Results and Analysis

As mentioned in Sect. 3.2, GOP-weight and GOP-FB were computed based on lattice free MII models. For comparison, we also calculated baseline GOP scores, GOP1 and GOP2 based on conventional DNN-HMM models as described in Sect. 3.1.

The correlations between different automatic scores are shown in Table 2. GOP-weight and GOP-FB show significant improvements over baseline GOP1 and GOP2.

Table 2. Correlations between automatic scores and reference scores

GOP1	GOP2	GOP-weight	GOP-FB
0.72	0.75	0.82	0.81

We further investigate the performances of these scores on different groups of students with beginner, intermediate and advanced proficiency levels. As shown in Table 3, the proposed GOP-weight and GOP-FB outperform GOP1 and GOP2 on data

in every proficiency group. GOP-weight shows higher performance with advanced intermediate learners than GOP-FB while GOP-FB performs better with beginners. We used linear regression models to combine to combine GOP-FB and GOP-weight with leave-one-out cross verification and yielded the best performance of 0.85 over all the data.

Table 3. Correlations between automatic scores and reference scores with different groups of learners

Proficiency	GOP1	GOP2	GOP-weight	GOP-FB
Beginner	0.51	0.58	0.62	**0.65**
Intermediate	0.43	0.49	**0.54**	0.52
Advanced	0.42	0.50	**0.59**	0.56

5 Conclusion

Two novel implementations of Goodness of Pronunciation (GOP) based on lattice free MMI were proposed to improve automatic scoring of L2 English speech. Experimental results show that by incorporating sequential information of speech, significant improvements have been found over the conventional GOP based automatic scoring methods based on DNN-HMM acoustic models within the Maximum Likelihood criterion. Future work includes combine different features to further improve robustness of automatic scoring.

References

1. Tsubota, Y., et al.: Practical use of english pronunciation system for Japanese students in the CALL classroom. In: Proceedings of ICSLP 2004, pp. 1689–1692 (2004)
2. Zhang, et al.: Generalized segment posterior probability for automatic Mandarin pronunciation evaluation. In: Proceedings of the ICASSP, pp. 201–204 (2007)
3. Neri, A., et al.: Automatic Speech Recognition for second language learning: How and why it actually works. In: Proceedings of International Congresses of Phonetic Sciences, pp. 1157–1160 (2003)
4. Cardenoso-Payo, V., et al.: Assessment of Non-native Prosody for Spanish as L2 using quantitative scores. In: Proceedings of LREC, pp. 3967–3972 (2014)
5. Luo, D., et al.: Automatic pronunciation evaluation of lan-guage learners' utterances generated through shadowing. In: Proceedings of the INTERSPEECH, pp. 2807–2810 (2008)
6. Dahl, G.E.,. et al.: Large-vocabulary continuous speech recognition with context-dependent DBN-HMMs. In: Proceedings of the ICASSP (2011)
7. Hu, W., et al.: A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL). In: Proceedings of the INTERSPEECH 2012, pp. 1886–1890 (2012)

8. Luo, D., et al.: Improvement of segmental mispronunciation detection with prior knowledge extracted from large L2 speech corpus. In: Proceedings of the INTERSPEECH 2011, pp. 1593–1596 (2011)
9. Luo, D., et al.: Naturalness judgement of L2 english through dubbing practice. In: Proceedings of the INTERSPEECH (2016)
10. Luo, D., et al.: Factorized deep neural network adaptation for automatic scoring of L2 speech in english speaking tests. In: Proceedings of INTERSPEECH 2018, pp. 1656–1660 (2018)
11. Witt, S.M., Young, S.J.: Phone-level pronunciation scoring and assessment for interactive language learning. Speech Commun. **30**(2–3), 95–108 (2000)
12. Dahl, G.E., et al.: Context-Dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEE Trans. Audio, Speech Lang. Process. **20**(1), 30–42 (2012)
13. Hadian, H., Sameti, H., Povey, D., et al.: End-to-end speech recognition using lattice-free MMI. In:. Conference of the International Speech Communication Association, pp. 12–16 (2018)
14. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of the ASRU (2011)
15. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210 (2015)