Variations of multi-task learning for spoken language assessment

Jeremy H. M. Wong, Huayun Zhang, and Nancy F. Chen

Institute for Infocomm Research, Singapore

{jeremy_wong, zhang_huayun, nfychen}@i2r.a-star.edu.sg

Abstract

Automatic spoken language assessment often operates within a regime whereby only a limited quantity of training data is available. In other low-resourced tasks, such as in speech recognition, multi-task learning has previously been investigated as a potential approach to regularise the model and maximise the utilisation of the available annotation information during training. This paper applies multi-task learning to spoken language assessment, by assessing three different forms of task diversities. These are, concurrently learning scores at different linguistic levels, different types of scores, and different representations of the same score. Experiments on the speechocean762 dataset suggest that jointly learning from phone and word-level scores yields significant performance gains for the sentence-level score prediction task, and jointly learning from different score types can also be mutually beneficial.

Index Terms: Multi-task learning, language assessment, low-resource, combination, diversity

1. Introduction

Spoken Language Assessment (SLA) aims to provide an automated evaluation of oral proficiency, and is an integral component within computer-aided language learning. The goal is to compute a proficiency score that is similar to one that an expert human rater would have assigned. Training Datasets available for SLA, such as [1], are often limited in size. Collecting and manually annotating larger quantities of diverse data may be expensive. Such low-resource tasks may have limited coverage between their training set and intended evaluation domain, making it challenging to ensure the robustness of the trained models. Improving the model performance, without having to resort to expensive data collection, may allow for more reliable and widespread use of automatic SLA. This may in turn permit language education to be more accessible to a wider population.

The model's robustness may be improved by expanding the coverage of the training set, either by generating additional data through simulation [2] or by augmenting existing data through perturbation [3]. Using external datasets can also increase the training data coverage, for example, in transfer learning from a different language [4]. Another solution that can aid in a low-resource scenario is multi-task learning [5], which trains a single model by simultaneously optimising multiple criteria. This is often motivated as being able to improve generalisation, because the hidden representations that the model learns to extract may be less specific to any single task [5]. Another motivation is that each task may provide a different form of information for the model to learn, thereby widening the diversity of information that the model can extract from the data.

When data is annotated with multiple label types, a model may learn from these jointly in a multi-task framework. In SLA, data can be annotated as different types of assessed scores and at different linguistic levels [1]. This paper investigates the effectiveness of using different varieties of annotation types to learn from jointly in a multi-task fashion in SLA. Three variations of task diversities are explored, namely different linguistic levels, different score types, and different score representations. Furthermore, the multiple outputs of a multi-task model can be interpreted as an ensemble [6]. This paper, therefore, also investigates applying multi-task ensemble combination to SLA.

2. Relation to prior work

Multi-task learning has been applied across different machine learning tasks, including natural language processing [7] and computer vision [8]. In Automatic Speech Recognition (ASR), previous studies have considered the secondary tasks of recognising multiple languages [9, 10], monophones and surrounding contexts [11], Mandarin tones [12], graphemes in addition to phonemes [13], and multiple sets of senones that are defined from an ensemble of different clustering trees [14]. Similarly to this paper, several of these works [11, 12, 13, 14] may also be interpreted as jointly learning from different annotation types. At the intersection of ASR and SLA, work in [15] trains an acoustic model jointly toward the separate tasks of adult and children ASR, then uses the model for mispronunciation detection. In the related field of speech emotion recognition, data can be annotated with the emotion class, together with attributes such as valence, activation, and dominance [16], and work in [17] studies the possibility of jointly learning from a variety of these annotation types. In this paper, multi-task learning is applied to SLA. The SLA model in [18] jointly learns to predict the pronunciation and fluency scores, but the effectiveness of such multi-task learning has not been explicitly proven, as no comparison against training toward only a single score type has been presented. Such a comparison is presented in this paper.

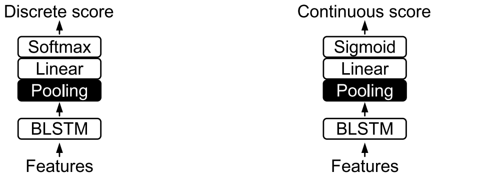
In Multi-task learning, diverse forms of annotation information may aid the model during training. During evaluation, ensemble combination can leverage upon the diversity of predictive behaviours. Works in [19, 20] combine diverse feature representations as inputs to a single neural network SLA model. In [21], multiple predicted scores are combined, using combination weights that are trained as a support vector machine. This paper uses the multiple outputs of a single multi-task model as an ensemble, as is done in [6] for ASR. Furthermore, the scores are combined as an equally-weighted average, avoiding the need to train the combination.

3. Forms of multi-task learning

A model can be trained toward multiple concurrent tasks, by optimising a weighted linear average between all of the tasks,

$$\mathcal{L}_{\text{multi-task}} = \sum_{m=1}^{M} \lambda_m \mathcal{L}_m, \tag{1}$$

This work was supported by the A⋆STAR Computational Resource Centre through the use of its high performance computing facilities.



(a) *Discrete categorical output* (b) *Contiuous scalar output*

Figure 1: *Single-output model topologies for SLA. Pooling is absent for a phone-level output*

where $\mathcal{L}_m$ is the criterion for the mth task, λ_m is its respective interpolation weight, and M is the number of tasks. It is often argued that the tasks used should be sufficiently different from each other, to encourage the learned hidden representations toward not being too specific to any one task [5]. At the same time, it may be useful for the tasks to be sufficiently similar to each other, to allow the information from the annotations of one task to be useful toward the prediction of another task. This consideration is particularly relevant if the model is operating near an under-fitting regime, as it may be counterproductive to waste the model's limited capacity on an irrelevant task. A balance between these competing considerations must therefore be sought among the choices of possible combinations of tasks.

3.1. Diversity of linguistic levels

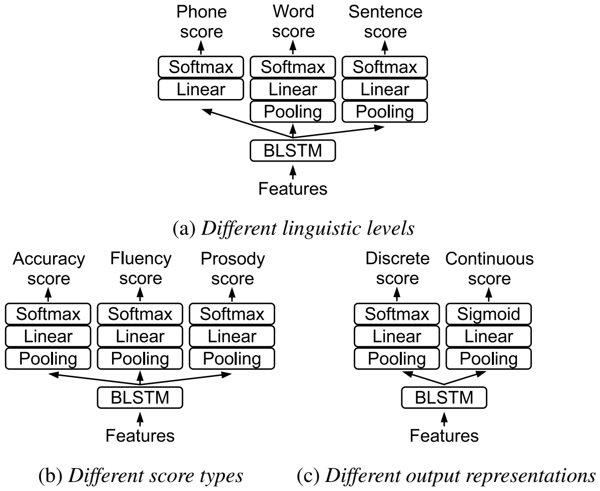
One possible set of tasks is to learn to predict the assessment scores at different linguistic levels [22]. An SLA dataset may be annotated with reference scores at the sentence, word, and phone levels. It may be reasonable to assume that a human rater's judgement of the oral proficiency at one linguistic level is correlated with that human rater's judgement of the oral proficiency at another level. Thus, the information obtained by learning to predict the score at one level may be helpful toward another level too. At the same time, a human rater may use different criteria to judge the scores at different linguistic levels. Therefore, the annotations from each level should still contain independent information.

3.2. Diversity of score types

This paper also assesses multi-task learning between different score types at the same linguistic level [22]. SLA datasets may be annotated with different types of scores. For example, the Malay and Tamil data used in [19] are annotated with the sentence-level pronunciation, fluency, and intonation, while data in [1] is annotated at the sentence level with accuracy, completeness, fluency, and prosody. According to [1], accuracy judges the pronunciation, while completeness judges the fraction of words within the sentence that are pronounced well. Fluency and prosody each judge different aspects of the speaking rate, pauses, repetitions, and stammering. Each score type assesses a different aspect of the oral proficiency, and at the same time, a speaker's proficiency as assessed by one score may be fairly correlated with the same speaker's proficiency as assessed by another score. Thus the tasks are related, but still diverse.

3.3. Diversity of output representations

The representation of the model's output score can also take diverse forms. In SLA, the model's output can represent either a categorical distribution over a set of discrete integer scores [18] or a continuous scalar variable [19]. Furthermore, different forms of training criteria may be preferred for each of these representations. For example, work in [18] trains a categorical output using a Cross-Entropy (CE) criterion, while [19] trains a scalar output using a Mean Squared Error (MSE) criterion. When using different output representations of the same score as the set of tasks, the different tasks are highly similar to each other. Diversity of the annotation information between the tasks may arise as different emphases placed by the different training criteria, or from the floating point information that is contained within the continuous scalar score annotation but is lost in the rounding of the discrete score annotation. Yet another difference may arise, because the scalar output captures the continuity of the score space, while the categorical output makes no assumption about the relative distances between the different score classes. This paper investigates training a model that uses multiple different output representations of the same score.



(a) *Different linguistic levels*

(b) *Different score types* (c) *Different output representations*

Figure 2: *Multi-task model topologies for SLA*

4. Model topology

The experiments used models with single outputs, and multi-task models with multiple outputs. For each sentence, the input was a sequence of feature vectors, with the sequence length corresponding to the number of phones in the sentence. Models with a single output are shown in figure 1 and are inspired by [19]. The features were fed into a Bidirectional Long Short-Term Memory (BLSTM) layer [23] that had 32 cells per direction. For models with word and sentence-level outputs, equally weighted mean pooling was applied to the BLSTM output, across all phones within each word or sentence. No pooling was used for a phone-level output, as the input and output have equal lengths. A linear layer was then used to form the final output dimension. Two forms of model outputs were considered. The output could represent a categorical distribution over integer scores, by using a softmax output layer, shown in figure 1a. Alternatively, a one-dimensional sigmoid output layer was used to compute a continuous scalar score, shown in figure 1b.

The multi-task models considered in this paper used a separate output for each task. The BLSTM layer was shared across all tasks and its output was branched into each of the tasks. The topologies for the multi-task models that use different linguistic levels, score types, and output representations are illustrated in figures 2a, 2b, and 2c respectively. It is hoped that information from the annotations of the multiple tasks can all contribute toward the training of the shared BLSTM layer, and that the representations computed by the BLSTM may be more general.

5. Score combination

A multi-task model can have multiple different outputs. In certain situations, it may be possible to treat each output as a member of an ensemble [6], and the predictions of these members can be combined together to leverage upon their different error patterns [24]. One such situation is when the separate tasks are to learn different output representations of the same score. The combination method used in this paper is a weighted average of the hypothesised scores from each ensemble member,

$$s_{\text{combine}} = \sum_{m=1}^{M} \eta_m s_m, \tag{2}$$

where s_{combine} is the combined score, s_m is the hypothesised score from the mth member in the ensemble, that has a combination weight of η_m, such that $\sum_m \eta_m = 1$ and $\eta_m \geq 0$. The advantage of this combination method over majority voting [24] is that it can be used when the ensemble has only two members. The advantage of this over probabilistic approaches [25] is that there is no requirement for the members of the ensemble to have outputs that must represent probability distributions.

6. Experiments

6.1. Setup

Experiments were performed on the speechocean762 dataset [1]. This comprises train and test sets with 2500 sentences each. 250 sentences from within the 2500 of the train set were held out for validation, without enforcing speaker set disjointment. The sentences were in English and were uttered as read speech by a disjoint group of 125 native Mandarin speakers in each of the train and test sets. Proficiency scores at the sentence, word, and phone levels were provided by five human raters. The five scores were averaged at the phone level, and the median was taken at the sentence and word levels to compute the reference score. The scores at the sentence and word levels range between 0 and 10, while the phone-level score ranges between 0 and 2. Following the terminology in [1], the sentence-level score types are accuracy, completeness, fluency, and prosody. The word scores are accuracy and stress. The phone score is accuracy.

As is done in the baseline recipe in [1], the experiments first used the Kaldi toolkit [26] to train a CE hybrid ASR model on the Librispeech dataset [27]. This was used to force align the train and test speechocean762 data. Following the setup in [19], goodness of pronunciation [28], log phone posterior [28], log posterior ratio [28], tempo [19], phonetic embedding [19], and pitch features were computed per phone, according to the forced alignment. The tempo features were spliced with left and right contexts of two phones each. All features were concatenated together to form a total dimension of 131. Global per-dimension mean and variance were computed across the train set, and used to normalise the features.

The SLA models, described in section 4, used a sequence of these feature vectors per sentence as input, with there being one feature vector per phone. Dropout layers [29], which zeroed out 60% of the dimensions, were placed before the BLSTM and linear layers. The SLA models were trained using an AdamW optimiser [30] with a fixed base learning rate.

Table 1: *Jointly learning multiple linguistic levels*

Training level	PCC / MSE for accuracy score at level phone	word	sentence
phone	0.486 / 0.12	-	-
word	-	0.544 / 2.29	-
sentence	-	-	0.672 / 1.37
multi-task	0.481 / 0.12	0.548 / 2.29	0.690 / 1.29

A model with a categorical output was decoded by choosing the expected score from the posterior distribution as the hypothesis. At the phone level, this decoding method may match well with the method of obtaining the reference scores, which was by averaging the scores across the multiple human raters. For a model with a continuous scalar output, the sigmoid output was taken as the decoded hypothesis, and was then scaled to the bounds of the score range. The SLA models were assessed by computing the Pearson's Correlation Coefficient (PCC) and MSE metrics, against the average or median human rater reference. Following [1], both the hypothesised and reference scores were rounded to the closest integers before computing the performance metrics.

6.2. Statistical significance

Statistical significance between two performance metrics was measured only at the sentence-level, because at the word and phone levels, it may be more questionable as to whether outputs of different words or phones within the same sentence are truly independent of each other. For MSE, the two-sided paired t-test was used. Unlike the MSE, the PCC cannot be easily expressed as an expectation over data samples. Thus, the central limit theorem cannot naively be used. Instead, by following the paired Z_1^* test approach in [31], an approximately normally distributed transformation [32] was first computed from the two PCCs being compared. Then, the significance was computed as the two-sided cumulative density of the transformed quantity.

6.3. Linguistic levels

The first experiment investigates using different linguistic levels to define the tasks. Separate single-output models were first trained at each of the sentence, word, and phone levels, toward their respective accuracies. All models had categorical outputs and were trained using the CE criterion. A multi-task model, shown in figure 2a, was also trained jointly toward all of the sentence, word, and phone accuracies. The multi-task training criterion, as in (1), was a weighted sum between the cross-entropies of the sentence, word, and phone-level accuracies. The interpolation weights were tuned, such that the CE curves on the validation set for each of the constituent criteria reached their respective valleys at approximately the same training iteration. When performing evaluation on the test set, only a single output of the multi-task model was used, with the choice of output corresponding to the linguistic level being evaluated. As such, the effective number of parameters in the multi-task model is the same as that in the single-output model during evaluation. As a performance baseline, a support vector machine approach for predicting the phone-level accuracy in [1] achieves PCC and MSE values of 0.45 and 0.16 respectively.

Table 1 compares the separate models against the multi-task model. The results show that multi-task learning yields gains at the sentence level, with significances of $\rho_{\text{PCC}} = 0.0006$ and

Table 2: *Jointly learning multiple score types*

Training score	PCC / MSE for sentence score of type accuracy	fluency	prosody
accuracy	0.672 / 1.37	-	-
fluency	-	0.715 / 1.02	-
prosody	-	-	0.725 / 0.99
multi-task	0.694 / 1.31	0.730 / 0.98	0.737 / 0.96

$\rho_{\text{MSE}} = 0.0019$. At the word and phone levels, the multi-task model performs similarly to each of the single-output models. This may suggest that information about language assessment at the finer linguistic levels may be useful toward the task of language assessment at the sentence level. A possible alternative explanation of this trend may be that training data is fairly sparse at the sentence level, with only one output reference label per sentence. Thus, sentence-level modelling may benefit more from the regularisation effect of multi-task learning than the finer linguistic levels do. To assess this, the sentence-level PCCs, measured on the 2500 sentences in the train set, are 0.799 and 0.836 for the single-output and multi-task models respectively. The differences between the train and test set PCCs are 0.127 and 0.146 for the single-output and multi-task models. Therefore, multi-task learning seems to widen the performance gap between the train and test sets, which does not support the hypothesis that multi-task learning is regularising the sentence-level behaviour. Despite this, the multi-task model still exhibits a better test set performance at the sentence level.

6.4. Score types

The next experiment investigates multi-task learning between different score types. All models used sentence-level categorical outputs. Single-output models were trained using CE criteria toward each of the sentence score types, namely the accuracy, fluency, and prosody. The completeness score was avoided here, because it has a fairly severe label imbalance that makes it difficult to use. A multi-task model with three outputs was trained toward the accuracy, fluency, and prosody scores, shown in figure 2b. Equal interpolation weights were used here.

Table 2 compares these single-output and multi-task models over the various sentence-level score types. Performance gains for the accuracy, fluency, and prosody are observed in the multi-task model, compared to each of the single-output models. The significances between the accuracies are $\rho_{\text{PCC}} = 0.0001$ and $\rho_{\text{MSE}} = 0.0227$, between the fluencies are $\rho_{\text{PCC}} = 0.0029$ and $\rho_{\text{MSE}} = 0.0272$, and between the prosodies are $\rho_{\text{PCC}} = 0.0389$ and $\rho_{\text{MSE}} = 0.1776$. The MSE performance gains are not always significant. The train set reference annotations for the three score types, after taking the median over the multiple raters, are themselves fairly correlated with each other, with PCCs of 0.805 between accuracy and fluency, 0.797 between accuracy and prosody, and 0.932 between fluency and prosody. Despite these high correlations, the results suggest that there may still be independent information within each separate score type that helps in the prediction of the other score types. This validates the use of such multi-task learning in [18].

6.5. Output representations

The final experiment assesses multi-task learning between different output representations of the same score type. All models were designed to compute the sentence-level accuracy. Two separate single-output models were trained. One used a continuous scalar output that was trained using an MSE criterion. The other used a categorical output that was trained using a CE criterion. A multi-task model with both of these output types, shown in figure 2c, was also trained. The multi-task interpolation weights were again tuned such that the constituent criteria computed on the validation set converged to valleys at about the same training iteration.

Table 3: *Jointly learning multiple output representations*

Training representation	PCC / MSE for output representation scalar	categorical	combine
scalar	0.690 / 1.41	-	0.695 / 1.27
categorical	-	0.672 / 1.37	
multi-task	0.670 / 1.57	0.669 / 1.40	0.672 / 1.39

The comparison between these single-output and multi-task models in table 3 shows that the multi-task model does not improve upon the performances of each of the single-output models. This suggests that it may not be beneficial to concurrently learn different representations of the same score.

In [6], ensemble combination is performed across the multiple outputs of a multi-task model. It is possible to apply combination to the ensemble with different output representations of the same score in table 3, to leverage upon the diversity of error patterns. The combination method of (2) was used, with equal combination weights. The results show slight but consistent gains yielded by the combinations, compared to either the separate single-output models or each of the separate multi-task outputs. The significances between the best single-output models of each metric and their combination are $\rho_{\text{PCC}} = 0.3992$ and $\rho_{\text{MSE}} = 3 \times 10^{-5}$. Thus, the combination gains do not exhibit consistent statistical significance. The combination gains for the multi-task model are even smaller. Some complementary diversity may exist between the error patterns arising from each of the different output representations, but the limited significance restricts the confidence in this conclusion.

Inspired by [33], the diversity between the hypotheses of two models can be assessed by computing the inter-model evaluation metrics, with one of the sets of hypotheses being treated as the reference. The PCC and MSE computed between the two sets of hypotheses from the multi-task outputs are 0.944 and 0.29 respectively, while those between the hypotheses of the two single-output models are 0.820 and 0.72. This shows that the diversity between the hypotheses of the two separate single-output models is wider than that between the two outputs of the multi-task model. This is despite both of the single-output models being in closer agreement with the reference, compared to the agreement between the multi-task hypotheses and the reference. The reduced diversity between the multi-task hypotheses may be due to the shared BLSTM layer and the joint training. This limited diversity may be a reason why the multi-task combination does not appear to yield any large performance gain.

7. Conclusion

This paper investigates the efficacy of various forms of multi-task learning in SLA, by considering different diversities of tasks. The results suggest that concurrently learning different linguistic levels and different score types may be beneficial, but learning different representations of the same score type may not be. A future extension of this work may consider jointly learning from all forms of task diversities at the same time.

8. References

[1] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, “speechocean762: an open-source non-native english speech corpus for pronounciation assessment,” in *Interspeech*, Brno, Czechia, Aug 2021, pp. 3710–3714.

[2] Y. Huang, L. He, W. Wei, W. Gale, J. Li, and Y. Gong, “Using personalized speech synthesis and neural language generator for rapid speaker adaptation,” in *ICASSP*, Barcelona, Spain, May 2020, pp. 7399–7403.

[3] N. Jaitly and G. E. Hinton, “Vocal tract length perturbation (VTLP) improves speech recognition,” in *ICML*, Atlanta, USA, June 2013.

[4] T. Schultz and A. Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition,” *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, Aug 2001.

[5] R. Caruana, “Multitask learning,” *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul 1997.

[6] O. Siohan and D. Rybach, “Multitask learning and system combination for automatic speech recognition,” in *ASRU*, Scottsdale, USA, Dec 2015, pp. 589–595.

[7] R. Collobert and J. Weston, “A unified architecture for natural language processing: deep neural networks with multitask learning,” in *ICML*, Helsinki, Finland, Jul 2008, pp. 160–167.

[8] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in *ECCV*, Zurich, Switzerland, Sep 2014, pp. 94–108.

[9] A. Mohan and R. Rose, “Multi-lingual speech recognition with low-rank multi-task deep neural networks,” in *ICASSP*, Brisbane, Australia, Apr 2015, pp. 4994–4998.

[10] V. H. Do, N. F. Chen, B. P. Lim, and M. A. Hasegawa-Johnson, “Multitask learning for phone recognition of underresourced languages using mismatched transcription,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 501–514, Mar 2018.

[11] M. L. Seltzer and J. Droppo, “Multi-task learning in deep neural networks for improved phoneme recognition,” in *ICASSP*, Vancouver, Canada, May 2013, pp. 6965–6969.

[12] J. Lin, W. Li, Y. Gao, Y. Xie, N. F. Chen, S. M. Siniscalchi, J. Zhang, and C.-H. Lee, “Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features using extended recognition networks,” *Journal of Signal Processing Systems*, vol. 90, pp. 1077–1087, Feb 2018.

[13] D. Chen, B. Mak, C.-C. Leung, and S. Sivadas, “Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition,” in *ICASSP*, Florence, Italy, May 2014, pp. 5592–5596.

[14] P. Bell, P. Swietojanski, and S. Renals, “Multitask learning of context-dependent targets in deep neural network acoustic models,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 2, pp. 238–247, Feb 2017.

[15] R. Tong, N. F. Chen, and B. Ma, “Multi-task learning for mispronunciation detection on Singapore children’s Mandarin speech,” in *Interspeech*, Stockholm, Sweden, Aug 2017, pp. 2193–2197.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemazadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, Nov 2008.

[17] R. Xia and Y. Liu, “A multi-task learning framework for emotion recognition using 2D continuous space,” *IEEE Trans. Affective Comput.*, vol. 8, no. 1, pp. 3–14, Jan 2017.

[18] R. Duan and N. F. Chen, “Unsupervised feature adaptation using adversarial multi-task training for automatic evaluation of children’s speech,” in *Interspeech*, Shanghai, China, Oct 2020, pp. 3037–3041.

[19] H. Zhang, K. Shi, and N. F. Chen, “Multilingual speech evaluation: case studies on English, Malay and Tamil,” in *Interspeech*, Brno, Czechia, Aug 2021, pp. 4443–4447.

[20] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, “Improving mispronunciation detection for non-native learners with multisource information and LSTM-based deep models,” in *Interspeech*, Stockholm, Sweden, Aug 2017, pp. 2759–2763.

[21] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, “Landmark-based automated pronunciation error detection,” in *Interspeech*, Chiba, Japan, Sep 2010, pp. 614–617.

[22] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, “Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment,” in *ICASSP*, Singapore, May 2022, pp. 7262–7266.

[23] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM networks,” in *IJCNN*, Montreal, Canada, Jul 2005, pp. 2047–2052.

[24] J. G. Fiscus, “A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER),” in *ASRU*, Santa Barbara, USA, Dec 1997, pp. 347–354.

[25] H. Xu, D. Povey, L. Mangu, and J. Zhu, “Minimum Bayes risk decoding and system combination based on a recursion for edit distance,” *Computer Speech and Language*, vol. 25, no. 4, pp. 802–828, Oct 2011.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, “The Kaldi speech recognition toolkit,” in *ASRU*, Hawaii, USA, Dec 2011.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *ICASSP*, Brisbane, Australia, Apr 2015, pp. 5206–5210.

[28] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, “Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers,” *Speech Communication*, vol. 67, pp. 154–166, Mar 2015.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, Jun 2014.

[30] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *ICLR*, New Orleans, USA, May 2019.

[31] J. H. Steiger, “Tests for comparing elements of a correlation matrix,” *Psychological Bulletin*, vol. 87, no. 2, pp. 245–251, 1980.

[32] O. J. Dunn and V. A. Clark, “Correlation coefficients measured on the same individuals,” *Journal of the American Statistical Association*, vol. 64, no. 325, pp. 366–377, Mar 1969.

[33] J. H. M. Wong and M. J. F. Gales, “Multi-task ensembles with teacher-student training,” in *ASRU*, Okinawa, Japan, Dec 2017, pp. 84–90.