Automatic Assessment Of Spoken English Proficiency Based On Multimodal & Multitask Transformers

Kamel Nebhi, Gyorgy Szaszak
Education First
Selnaustrasse 30
8001 Zürich, Switzerland
{kamel.nebhi, szaszak.gyorgy}@gmail.com

Abstract

This paper describes technology developed to automatically grade students on their English spontaneous spoken language proficiency with common european framework of reference for languages (CEFR) level. Our automated assessment system contains two tasks: *elicited imitation* and *spontaneous speech assessment*. Spontaneous speech assessment is a challenging task that requires evaluating various aspects of speech quality, content, and coherence. In this paper, we propose a multimodal and multitask transformer model that leverages both audio and text features to perform three tasks: scoring, coherence modeling, and prompt relevancy scoring. Our model uses a fusion of multiple features and multiple modality attention to capture the interactions between audio and text modalities and learn from different sources of information.

1 Introduction

Language proficiency testing is an increasingly important part of our society. The need to demonstrate language abilities through standardized testing is required in many situations for access to higher education and employment opportunities.

This paper presents an automatic system to address the assessment of English spoken proficiency with CEFR level. Our framework contains two tasks: *elicited imitation* and *spontaneous speech assessment*.

The *elicited imitation* task taps into reading and speaking skills by requiring examinees to say a sentence out loud. Test takers must be able to process the input and are evaluated on their fluency, accuracy, and ability to use complex language orally (Van Moere, 2012). We employ statistical machine learning (ML) and natural language processing (NLP) using a transformer-based classifier to directly estimate item difficulties for a large item bank.

For *spontaneous speech assessment*, the candidates are asked to talk about a prompt/question-related topic. Our spontaneous speech system is based on EF Standard English Test (EFSET) dataset. In the proposed system, the students' spoken answers are first transcribed by a state-of-the-art automatic speech recognition (ASR) system and then scored using a multimodal and multitask framework. This work argues that audio and text features are complementary for a valid automatic spoken assessment system (Mayfield and Black, 2020; Gretter et al., 2019).

The contributions in this paper are threefold: 1) we propose the use of test items for elicited imitation that can be automatically created and graded using a BERT transformer; 2) a multimodal and multitask framework for spontaneous speech assessment combining audio and text is proposed; 3) a complete automated assessment framework was built and evaluated using a calibrated dataset.

In the pages that follow, we first summarize the state-of-the-art in automated speech assessment and then describe our approach to assess language proficiency. We then present evidence for the validity and reliability of our approach using EFSET validation set and a calibration dataset. Finally, we will give a conclusion.

2 Related Works

A number of approaches have been proposed to assess different aspects of a learner's spoken language proficiency. Most automatic assessment systems contain an ASR system, with the success of deep neural networks (DNN) in speech recognition (Hinton et al., 2012), a number of automatic assessment systems that deploy DNN-based speech recognition systems have been proposed. The extracted features are then used to train a grader to give a score. All existing automatic assessment sys-

Proceedings of Recent Advances in Natural Language Processing, pages 769–776
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_083

tems are learning-based and can be classified based on whether they are feature-based, end-to-end or multitask approaches.

2.1 Features-based approach

The Educational Testing Service (ETS) presented an automatic assessment system focused on spontaneous speech, named SpeechRater (Higgins et al., 2011; Zechner et al., 2009). SpeechRater exploits features related to pronunciation (audio and fluency features), grammatical accuracy and ASR confidence. This system gives a correlation of 0.7 with human scores on a dataset from the Test of English as a Foreign Language (TOEFL).

In Wang et al. (2018), an automatic assessment system for spontaneous speech of English is proposed using data from the Business Language Testing Service (BULATS) Online Speaking Test of Cambridge English Language Assessment. This system uses a deep neural network ASR system to generate transcriptions from which a set of features are extracted. In addition to audio and fluency features, they also exploit confidence, syntactic parsing (Briscoe, 2006) and pronunciation features. This system shows a Pearson Correlation Coefficient (PCC) of 0.865 and Mean Squared Error (MSE) of 10.2 when compared with expert scores.

Gretter et al. (2019) introduced an automatic assessment system using a DNN ASR system and then scored students' answers using a feedforward neural network that processes features extracted from the automatic transcriptions. In addition to audio signals, the system uses a set of LMs trained over different types of text data to compute features. The system was trained using the Trentino evaluation campaigns on trilinguism. This system shows a correlation of 0.7 and a weighted kappa of 0.77 when compared with expert scores.

Recently, Bamdev et al. (2023) presents a machine learning-based approach to assess the English proficiency of non-native speakers from their speech samples. The paper uses the SLTI SOPI dataset, which contains 1200 speech samples with different proficiency levels, rated by human experts on a scale from 1 to 5. The paper extracts various linguistic features from the speech samples, such as pronunciation, fluency, vocabulary, grammar, and discourse. They train two types of machine learning models to predict the proficiency scores from the linguistic features: a classification model that assigns each speech sample to one of the five proficiency levels using support vector machines (SVMs), and a regression model that outputs a continuous score between 1 and 5 using random forest regressors (RFRs). The paper reports that the regression model achieves a higher accuracy of 0.82 than the classification model with 0.77, based on the correlation with human scores.

2.2 End-to-End approach

Chen et al. (2018) proposed an end-to-end approach based on bidirectional long short-term memory (BD-LSTM) using attention mechanism and regression. This system performs better than the initial SpeechRater framework developed by ETS. The conventional model shows a PCC of 0.58 when the end-to-end approach provides higher performance with 0.60.

Grover et al. (2020) proposed a multi-modal end-to-end neural approach for automated assessment of non-native English speakers' spontaneous speech using attention fusion. The pipeline employs BD-RNN and BD-LSTM neural networks to learn complex interactions among acoustics and lexical features. They used data collected by Second Language Testing Inc. (SLTI) administrating Simulated Oral Proficiency Interview (SOPI) for L2 English speakers. The model shows a weighted kappa of 0.50 and 0.32 of MSE.

Recently, Singla et al. (2021) introduces a speaker-conditioned hierarchical model that assesses the language proficiency of speakers based on their oral responses. The model leverages a two-level attention mechanism to relate the prompts and responses, and speaker embeddings to capture individual variations. The model outperforms the baselines on human-machine agreement and provides insights into the learned representations. The paper shows that the model attains an average QWK of 0.82 on four datasets, which is a 6.92% increase over the baseline model.

2.3 Multitask Approach

Muangkammuen and Fukumoto (2020) presents a multi-task learning model that combines automated essay scoring and sentiment analysis. The model uses a hierarchical neural network to predict a holistic score and sentiment classes at different levels of text. The paper shows that sentiment features can improve essay scoring for some prompts.

More recently, Yang et al. (2022) proposes a multi-task learning framework that incorporates relevance and coherence modeling as auxiliary tasks

for automated text scoring. The paper uses negative sampling to generate samples for the auxiliary tasks and evaluates the model on the ASAP dataset. The paper reports that the model improves the QWK scores by 1.5% on average compared to other neural network models.

3 Proposed Approach

In this section, we are going to describe our system which combines *elicited imitation* and *spontaneous speech assessment*.

3.1 Elicited Imitation

The Elicited Imitation (EI) is a testing method that usually requires participants to listen to a series of stimulus sentences and then repeat the sentences as closely as possible. EI has been widely used as a measure of oral proficiency in second language acquisition research (Kostromitina and Plonsky, 2021; Wu et al., 2021).

Test takers must be able to process the input (e.g., orthography and grammatical structure) and are evaluated on their fluency, accuracy, and ability to use complex language orally (Van Moere, 2012). In practice, test items are written by experts. This labor-intensive process often restricts the number of items that can be created. To tackle this problem, we propose the use of test item formats that can be automatically created and graded using NLP.

3.1.1 Test Items Construction

To estimate item difficulty for the EI task, we employ statistical NLP to automatically project items onto a 3-point scale (elementary, intermediate, advanced).

These levels were assigned using an NLP model (sentence complexity classifier) trained on *newsinlevels* dataset. The *newsinlevels* corpus consists of 12,000 sentences ranked by 3 reading levels (elementary, intermediate, advanced).

Class	Precision	Recall	F-1 Score
Elementary	0.86	0.95	0.90
Intermediate	0.68	0.64	0.66
Advanced	0.86	0.81	0.83

Table 1: Performance of BERT-based Sentence Complexity Classifier.

We use a transformer-based architecture (BERT, (Devlin et al., 2018)) that has been pretrained on a large unlabeled corpus, and finetune it on *newsinlevels* dataset. Our model achieved 82% of accuracy on a validation dataset. Table 1 shows detailed performance of our BERT-based Sentence Complexity Classifier.

To build a bank of sentences, we downloaded 2000 English sentences from Tatoeba[1] (a free crowdsourced database of self-study resources for language learners). Then we apply our sentence complexity ranker to the Tatoeba dataset. Finally, we obtained a list of sentences annotated with the 3 difficulty levels.

To construct our final item list, we filtered the Tatoeba dataset with these features:

- length of sentence - 3 length bands: short (<8 syllables), medium (8-15 syllables), long (> 15 syllables) ;
- grammatically acceptable sentences: we selected acceptable English sentences from the grammar perspective ;
- non-profane sentences.

Table 2 shows examples of sentences, rated for predicted difficulty by the BERT complexity classifier model.

3.1.2 Automated Speech Scoring for Elicited Imitation

Our elicited imitation assessment method is based on local features derived from automatic speech recognition, e.g., the Goodness of Pronunciation (GOP) score. It takes the probabilities of the phonemes and processes them into the phoneme-level scores. In addition, it uses a process called "Forced Alignment" to align the targeted words and phonemes to the 10-millisecond audio frames from the given audio input.

3.2 Multimodal & Multi-task Learning for Spontaneous Speech Assessment

Our multimodal architecture consists of two parallel branches, the audio modality-based branch, and the text modality-based branch which consists of a multitask BERT model. Its core mechanisms are the fusion of multiple feature vectors and multiple modality attention.

From the audio data, we extract three kinds of features that belong to the audio modality: acoustic, prosodic, and spectral. A Time Delay Neural

[1] https://tatoeba.org

Candidate Sentence	Predicted Level
You are in my way.	Elementary
Humans were never meant to live forever.	Intermediate
I was wondering if you were going to show up today.	Advanced

Table 2: Example sentences, rated for predicted difficulty by the BERT complexity classifier model

Network (TDNN) then transforms these features into high-level representations.

We use a multitask BERT model to extract word embeddings from the text data that belong to the text modality. A fully connected layer then transforms these embeddings into contextual representations.

We concatenate the outputs of the TDNN and the fully connected layer to fuse multiple feature vectors. We apply a multi-head self-attention mechanism to the concatenated vector to fuse multiple-modality attention, which can model the interactions and relationships among different modalities and features. The model produces a CEFR score by a fully connected layer and a softmax layer as the final output.

Figure 1 shows the structure of the attention-based mechanism multimodal multitask model.

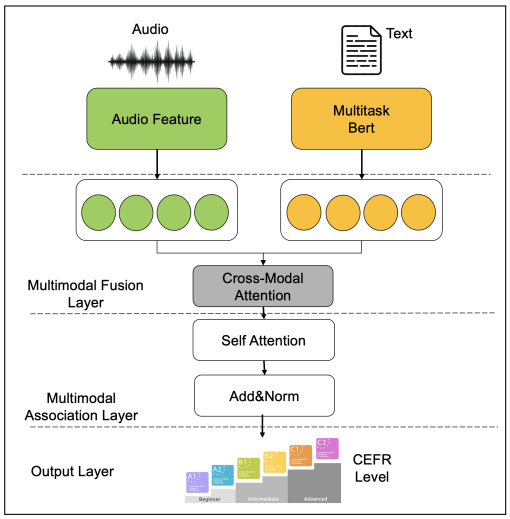

Figure 1: Structure of multimodal and multitask learning model.

3.2.1 TDNN model

Automatic Speech Recognition (ASR) is based on a hybrid Time Delay Neural Network (TDNN) acoustic models trained with the kaldi ASR toolkit on a mix of 9k hours of in-house data and LibriSpeech (Peddinti et al., 2015). For the scoring model, we restrict in house data for utterances with the best pronunciation scores. 3-fold speed perturbation is used to augment the training data. No augmentation with noise was used, although the in-house part of the dataset reflects various background conditions w.r.t. additive noise. We did not split the ASR training dataset w.r.t. native language or clustered it for accents, in order to make the resulting system simpler. As language model, an ARPA tri-gram is used for transcription with the transcription acoustic model in a single decoding pass.

Beside mel filterbank spectra, we also compute fundamental frequency contour directly from audio and silence/pause duration patterns as well as hesitation statistics from the alignment provided by the ASR during decoding. These supra-segmental features can be extracted quite reliably and can be used to assess intonation and stress patterns as well as fluency. The essential statistics for pause and hesitation include frequency of occurrence and duration (mean, standard deviation). Fundamental frequency can be used to assess intonation and stress patterns. We measured a Word Error Rate (WER) of 20.6% on elicited speech transcription on our in house 9 hours audio test set.

Phone quality is also influenced by stressing, in unstressed vowels reduction may take place. This can also be exploited in the assessment of proper stressing as part of fluency. The transcription acoustic models were created such that for most vowels, both a stressed and an unstressed variant is used and trained. In languages with lexical stress, such as English, this differentiation is simple and can be represented at the dictionary level.

Generally, the more hesitations are present, and the more and longer the pauses get, the least fluent is the speech, supposing we keep the expected speaking style constant. In tasks where speaking style is less formal, however, disfluencies such as hesitation and pauses are natural phenomena and hence, assessment is prevented from assigning lower fluency scores in such cases.

3.2.2 Multitask learning

Multi-task learning (MTL) is a machine learning technique that learns multiple tasks at the same time by sharing information among them (Crawshaw, 2020). MTL can improve the performance of each task compared to learning them separately. In MTL, there is a main task and some auxiliary tasks that can benefit from each other and enhance the generalization ability. The basic assumption for auxiliary tasks is that they should be relevant to the main task and help the main task learn better.

Discourse structure and coherence are important aspects of student answers and are often a part of grading rubrics. We describe the transformer-based discourse features that have been used to measure prompt relevancy and coherence.

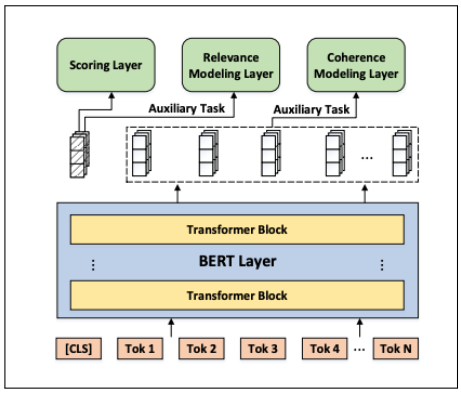

Figure 2: An overview of our multi-task learning architecture.

Scoring task is the main task of our model. It aims to predict a score for each essay. We employ a dense layer with a linear activation function to compute the score for each candidate answer based on the text representation R. The text representation R is a high-dimensional vector that encodes the semantic and syntactic information of both the prompt and the answer. We modify the output layer to produce a single scalar value and we use the mean squared error as a loss function.

$$y = W^T B(x) + b \tag{1}$$

where y is the predicted value, W is the weight vector of the output layer, $B(x)$ is the output of BERT for the input text x, and b is the bias term of the output layer.

Coherence modeling measures conceptual relations between different units within a response. Our approach measures overall coherence by calculating the semantic relatedness between adjacent sentences. Obviously, coherence scores for well-organized answers should be higher than the disorganized/random answers.

We use the BERT pre-trained language model (Devlin et al., 2018) and fine-tune it on EFSET dataset[2] using a fully connected perceptron layer. We leverage the Next Sentence Prediction objective of BERT and get a single representation for both sentences *s1* and *s2*. Given the sentence pair P_{ij}, the embedding of the [CLS] symbol from the top layer of BERT is denoted as C_{ij}. Owing to the Next Sentence Prediction pre-training objective of BERT, this vector C_{ij} is able to aggregate the semantic relations for the input sentence pair and is capable of identifying the relative order between two sentences. The softmax function is defined as:

$$Pij = softmax(WCij + b) \tag{2}$$

where W and b are the parameters of the fully connected perceptron layer, and Pij is the probability of sentence s_i preceding sentence s_j.

To find the right order of the sentences we use topological sort (Prabhumoye et al., 2020; Tarjan, 1976). Finally, we use the sentence accuracy metric (Logeswaran et al., 2018) to quantify the coherence of answers. Sentence accuracy measures the percentage of sentences for which their absolute position was correctly predicted.

Our model aims to reorganize an unordered set of sentences into a coherent paragraph. Then, the coherence score for well-organized answers should be higher than the incoherent answers.

Prompt-relevancy features measure how well the answer matches the prompt. We assume that the essay content and the topics are closely related. ATS systems may assign a high score to an essay that is well-written but off-topic. However, a human rater will prefer essays that are on-topic and penalize essays that are not. To capture the prompt-specific knowledge, we design an auxiliary task called prompt-relevancy modeling. We take the top 40% essays of all prompts and shuffle them, and use their prompts as labels. Then, we feed the latent text representation R learned from BERT into a dense layer with a softmax activation function to predict the prompt.

$$P = softmax(WR + b) \tag{3}$$

[2]We filtered the dataset using the coherence score provided by expert. Then we generated permuted sentence samples. Finally, we built a training set of 35000 samples and a test set of 9500 samples.

where P is the predicted prompt, W is the weight matrix, b is the bias vector, and $softmax$ is the activation function.

3.2.3 Multimodal Fusion Layer

The Multimodal Fusion Layer fuses multimodal data features.

In our approach, we use two main forms of multimodal sequence data: text (T) and audio (A). The modal features are extracted by different methods, which produce different dimensional features for text and audio sequences T $\in T, A$.

To align the sequences and make them have the same dimension, we apply 1D temporal convolutional layer as the final step.

Cross-modal Attention leverages the information exchange between text and audio modalities to fine-tune the weights of the model and the pre-trained language model BERT. The data processing layer produces the text features and audio features, respectively.

3.2.4 Multimodal Association Layer

The output sequence of the last layer of BERT encoder text is combined with the attention using residual connection and layer normalization (Add&Norm). This allows the network to stack more layers without suffering from vanishing gradients and also enhances the model accuracy and convergence rate.

The output sequence of the last layer of the BERT encoder for text for each task is combined with the attention weights from the cross-modal attention layer using residual concatenation, which adds the two sequences element-wise. Then, the resulting sequence is normalized using layer normalization, which scales and shifts the sequence to have zero mean and unit variance. This process of residual concatenation and normalization (Add&Norm) helps to stabilize the training and improve the performance of the multimodal architecture.

3.2.5 Output Layer

The last layer of our multimodal model is a softmax function that outputs a CEFR score between 1 and 6 for each input pair of audio and text. The softmax function is defined as:

$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \quad (4)$$

where x_i is the input to the function, which in our case is a linear combination of the concatenated features from the audio and text branches and n is the number of elements in the vector.

4 Experiments

4.1 Dataset

In this section, we present corpora that have been used to train and evaluate our system.

4.1.1 EF Standard English Test - Spontaneous Speech Assessment

The EF Standard English Test[3] (EFSET) dataset is based on a standardized test of the English language designed for non-native English speakers. EFSET contains around 4100 student tests (each test containing 14 prompts) annotated by teachers. Each student test is annotated with 4 scores between 0-100 representing accuracy, fluency, range and coherence. The 4 scores are then mapped to a final score using weights[4].

$finalscore = accuracy * 0.3 + fluency * 0.3 + range * 0.3 + coherence * 0.1$

4.1.2 EF Speak Oral English Test - Calibration dataset

For this experiment, we created a calibration/gold standard dataset to evaluate our experiments.

We used the online outsourcing platform Upwork to target English teachers or tutors and ask them to distribute the test to their students. Students could not submit the test twice and no additional instruction and information was given to pass the test. The test takers are from three continents: Africa (Nigeria), Europe (Albania, Ukraine, Turkey), and Asia (Philippines and Korea).

A total of 400 responses have been collected and totally 10 expert scorers participated in the scoring of the tests. The two parts of the tests are scored individually, and the scorer could not associate the parts as the information of students is anonymous. In the scoring process, a few individual audios are regarded as technical issues, which is defined as either the audio cannot be played or is inaudible. We remove the parts marked as technical issues and only reserve the test parts so that all the audio recordings are properly scored by the scorers. As a result, there are 379 test results and scores qualified.

[3] https://www.efset.org/

[4] These weights resulted from a calibration process that occurred during the test creation.

4.2 Evaluation

To evaluate our system, we use the Quadratic Weighted Kappa (QWK) and Pearson Correlation Coefficient (PCC). Table 3 shows the performance of our multimodal multitask framework compared to the expert graders for the EFSET test set. Our baseline system (multitask only) obtains a QWK score of 0.80 on the test set which shows a substantial agreement and a PCC of 0.8. When the system combines multimodal and multitask learning, it improves the QWK to 0.84 and the PCC to 0.86, showing a higher agreement and a stronger correlation.

To compare these results with recent works, (Singla et al., 2021) reports that their hierarchical model achieves an average QWK of 0.82 across four datasets, which is slightly lower our framework on EFSET. Another features-based approach provided by (Bamdev et al., 2023) reports that the system achieves a QWK of 0.81 on SLTI SOPI dataset, which is also lower than our model on EFSET. These papers suggest that the multimodal multitask framework has a competitive performance in automated speech scoring compared to other recent works.

Table 4 shows the performance of our framework on calibration evaluation set for EI and SSA tasks. Our system obtains 0.78 of QWK and 0.82 of PCC for both tasks. Figure 3 illustrates the associations between our test scores and IELTS. There is a strong correlation between our scores and IELTS.

Model	QWK	PCC
Multitask BERT (only)	0.80	0.83
Multitask BERT+Multimodal	0.84	0.86

Table 3: Performance of the Multimodal & Multitask framework compared to the expert graders.

Test Part	QWK	PCC
EI	0.71	0.79
SSA	0.84	0.86
EI+SSA	0.78	0.82

Table 4: Performance of the complete framework (EI and Spontaneous Speech Assessment) compared to the calibration dataset.

5 EF Speak Oral English Test

The EF Speak Oral English Test is an online assessment initially created using the methods in this paper. The elicited imitation task contains 9 items ranked by difficulty using our BERT classifier. The spontaneous speech assessment task contains 6 prompts. Figure 4 shows examples of items. Finally, each part is scored by our framework and the final value is mapped to the corresponding CEFR level.

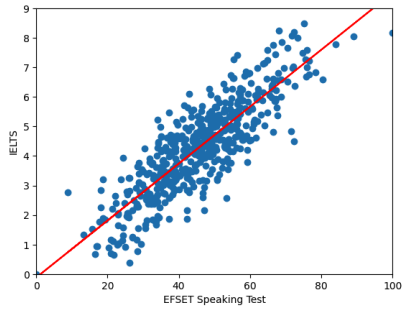

Figure 3: Relationship between EFSET Speaking Test scores and IELTS proficiency levels, as shown by scatterplot and pearson correlation coefficient ($r = 0.83$).

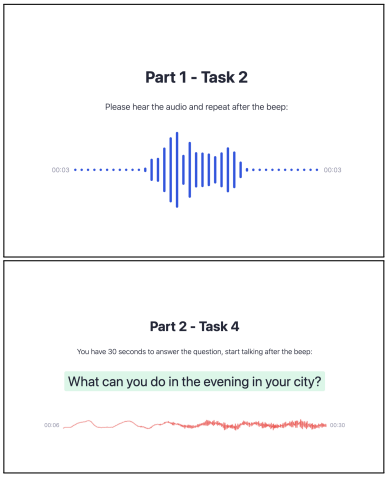

Figure 4: Example of test items for Spontaneous Speech Assessment.

6 Conclusion

This paper has described an automatic assessment system for spontaneous English based focused on *elicited imitation* and *spontaneous speech assessment*. This system uses a multimodal and multitask framework to leverage both audio and text features. The performance of the proposed system has been evaluated using PCC and QWK measures and the best combination of features gives a PCC of 0.86 and a QWK of 0.84 when compared with expert scores.

References

Pakhi Bamdev, Manraj Singh Grover, Yaman Kumar Singla, Payman Vafaee, Mika Hama, and Rajiv Ratn Shah. 2023. Automated speech scoring system under the lens: Evaluating and interpreting the linguistic cues for language proficiency. *International Journal of Artificial Intelligence in Education*, 33(1):119–154.

Ted Briscoe. 2006. An introduction to tag sequence grammars and the rasp system parser. Technical report, University of Cambridge, Computer Laboratory.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018. End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6234–6238. IEEE.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Roberto Gretter, Marco Matassoni, Katharina Allgaier, Svetlana Tchistiakova, and Daniele Falavigna. 2019. Automatic assessment of spoken language proficiency of non-native children. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7435–7439. IEEE.

Manraj Singh Grover, Yaman Kumar, Sumit Sarin, Payman Vafaee, Mika Hama, and Rajiv Ratn Shah. 2020. Multi-modal automated speech scoring using attention fusion. *arXiv preprint arXiv:2005.08182*.

Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, 25(2):282–306.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Maria Kostromitina and Luke Plonsky. 2021. Elicited imitation tasks as a measure of l2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, pages 1–26.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

Panitan Muangkammuen and Fumiyo Fukumoto. 2020. Multi-task learning for automated essay scoring with sentiment analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 116–123, Suzhou, China. Association for Computational Linguistics.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*.

Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2020. Topological sort for sentence ordering. *arXiv preprint arXiv:2005.00432*.

Yaman Kumar Singla, Avyakt Gupta, Shaurya Bagga, Changyou Chen, Balaji Krishnamurthy, and Rajiv Ratn Shah. 2021. Speaker-conditioned hierarchical modeling for automated speech scoring. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1681–1691.

Robert Endre Tarjan. 1976. Edge-disjoint spanning trees and depth-first search. *Acta Informatica*, 6(2):171–185.

Alistair Van Moere. 2012. A psycholinguistic approach to oral language assessment. *Language Testing*, 29(3):325–344.

Yu Wang, MJF Gales, Kate M Knill, Konstantinos Kyriakopoulos, Andrey Malinin, Rogier C van Dalen, and Mohammad Rashid. 2018. Towards automatic assessment of spontaneous spoken english. *Speech Communication*, 104:47–56.

Shu-Ling Wu, Yee Pin Tio, and Lourdes Ortega. 2021. Elicited imitation as a measure of l2 proficiency: New insights from a comparison of two l2 english parallel forms. *Studies in Second Language Acquisition*, pages 1–30.

Yupin Yang, Jiang Zhong, Chen Wang, and Qing Li. 2022. Exploring relevance and coherence for automated text scoring using multi-task learning. In *The 34th International Conference on Software Engineering and Knowledge Engineering, SEKE 2022, KSIR Virtual Conference Center, USA, July 1 - July 10, 2022*, pages 323–328. KSI Research Inc.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895.