# NEURAL APPROACHES TO AUTOMATED SPEECH SCORING OF MONOLOGUE AND DIALOGUE RESPONSES

*Yao Qian, Patrick Lange, Keelan Evanini, Robert Pugh, Rutuja Ubale, Matthew Mulholland, Xinhao Wang*

Educational Testing Service R&D, USA
{yqian, plange, kevanini, rpugh, rubale, mmulholland, xwang002}@ets.org

## ABSTRACT

We present Neural Network (NN) approaches to the automated assessment of non-native spontaneous speech in a monologic task and a simulated dialogic task. Three attention-based Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Networks (RNN) are employed to learn three dimensions (i.e., delivery, language use, and content) of scoring rubrics for the spoken responses. The prompts or turn history information are encoded to low-dimensional vectors by either a BLSTM-RNN or an end-to-end memory network (MemN2N) and used as the conditions of the inputs of the NN for rating the subscore of content. The three subscores are fused together to generate a holistic score. The experimental results show that our approaches significantly outperform the conventional approaches to speech scoring and the correlations of automatically predicted scores with the reference human scores are higher than human-human agreement levels for both tasks.

***Index Terms***— automated speech scoring, LSTM, RNN, attention, end-to-end memory networks

## 1. INTRODUCTION

Automated systems for scoring non-native speech assess spoken language proficiency along several dimensions of communicative competence including delivery (pronunciation, stress, fluency, and intonation), language use (vocabulary and grammar), content (topical relevance and appropriateness), and organization (discourse structure and coherence). It is an attractive but challenging application of spoken language technologies. ETS's SpeechRater$^{TM}$ [1,2] is one such scoring application and has been used to score open-ended, spontaneous responses to assessments of English for academic purposes. Each spoken response is first processed by speech processing technologies, where the input speech is transcribed into a sequence of linguistic units (phonemes, syllables, and words) by automatic speech recognition (ASR), and the corresponding features, which can be used to assess pronunciation, stress, fluency, and intonation, are extracted via forced-alignment with the recognized hypotheses. The recognized word sequence is then fed into a natural language processing module to generate the features related to vocabulary, grammar, content, and structure. All the features are then used to predict a score using a scoring model trained (in the sense of supervised learning) on responses scored by humans.

Recent advances in ASR and spoken language processing have led to improved systems for automated assessment of spoken language. Some lexical and syntactic features can be more accurately generated to address content appropriateness, topicality correctness, task completion, and pragmatic competence in some advanced automated speech scoring systems [3-10]. However, the features used in those advanced systems are still mostly handcrafted and aggregated (e.g. a bag of words), which cannot capture the nature of speech communication process, i.e., contextual information or temporal dynamics. In this paper, we use three BLSTM-RNNs for assessing spoken language proficiency in terms of delivery, language use and content, individually, and combine them together to predict a holistic score for a test taker's response. To incorporate dialog history (multi-turn interaction) for contextual information, end-to-end memory networks (MemN2N) [19] are explored to grade spoken dialogue responses.

## 2. RELATED WORK

Recently there have been several studies [11-14] on investigating neural networks for modeling sequential data (such as word sequences) for automated written essay scoring that show better performance compared to conventional approaches based on engineered and aggregated features. In [11], a hierarchical convolutional neural network (CNN) architecture was employed and competitive performance was shown for in-domain and domain-adaptation tasks. [12] found that a mean-over-time layer on top of an LSTM recurrent layer achieved the best performance among various neural network structures. A BLSTM-RNN with a weighted linear combination of two loss functions, score prediction and word embeddings, in multi-task learning was proposed in [13].

To date, only a few studies have been conducted using sequential features for automatically assessing spontaneous speech. Siamese Convolutional Neural Networks (CNN) and neural attention-based response-prompt relevance model have been used to detect off-topic responses in automated speech scoring systems [15,16]. In our previous studies, 1) the abstractions learned by BLSTM-RNN from ASR-free low-level time-sequence features, such as frame-level MFCC and F0, are jointly optimized with ASR-generated, aggregated features, such as the number of words per second, for predicting holistic scores of non-native spontaneous speech [17]; 2) two BLSTM-RNNs (one for the word embeddings of the recognized words, and other for word-level acoustic features, for example, word duration), were concatenated and fed into a linear regression layer for grading spoken responses [18]. The work presented in this paper focuses on investigating the different levels and the different contexts of sequential features for assessing the different dimensions of spoken language proficiency for both monologic and dialogic responses, which were not considered in the earlier studies. To the best of our knowledge, no work so far has studied using utterance history from the conversation on both sides for spoken language assessment.

## 3. DATA AND TASK

Two non-native spontaneous English corpora are used in this study. The first is drawn from the Listen Speak (LS) task of the TOEFL Junior Comprehensive assessment [20]. In this task type, the test

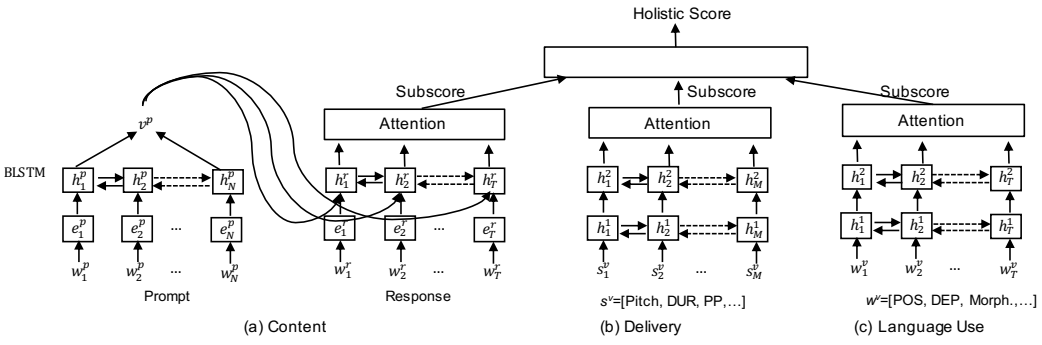


**Fig.1:** *The neural network architecture for scoring spoken responses*

taker listens to an audio stimulus (approximately 2 minutes in duration) containing information about a non-academic topic (for example, a class field trip) or an academic topic (for example, the life cycle of frogs) and provides a spoken response that should contain pieces of information that were provided in the stimulus. Each speaker in this corpus provided two or three responses to LS tasks. The responses are approximately 60 seconds in duration and contain roughly 100-150 words on average. Each response is scored on a scale of 0-4 following scoring rubrics on content, delivery, and language. This corpus is hereafter referred to as the monolog corpus.

The second corpus is also drawn from a pilot assessment of non-native English speaking proficiency for academic purposes. This assessment contained a task type in which the test taker is presented with a set of stimulus materials, such as a course schedule, an advertisement for a job on campus, etc., and is then presented with a series of spoken prompts from a computer-based interlocutor in the form of a simulated dialog. After each response from the test taker, the subsequent prompt from the computer-based interlocutor is played, until the final prompt has been reached. Expert human raters provided proficiency ratings on a scale of 0-5 for an entire simulated dialog based on the language learner's spoken English proficiency and how well the task was completed. The corpus and scoring rubrics are described in further details in [6]. This corpus is hereafter referred to as the dialog corpus.

## 4. NEURAL ARCHITECTURES FOR SCORING MODELS

To model the evolution of the response over time, we formulate automated speech scoring as a time-series regression problem of predicting a score for a sequence of features extracted from a given utterance. RNNs [21] configured to process input sequences of arbitrary length and capture temporal dynamics have been successfully applied to solve a wide range of machine learning problems with sequential data. With LSTM cells [22], RNNs can overcome the vanishing gradient problem when the input sequences are long. Gated recurrent units (GRU) [26] can be an alternative to solve the vanishing gradient problem but their performance is inferior to LSTMs in our tasks. Similar findings are reported in [12]. Recently, attention mechanisms have been shown to perform very well on many sequence-to-sequence mapping tasks such as speech recognition and machine translation [23,24]. The attention mechanism can be simply seen as a method for making the model focus on the states that are of high importance. Our previous work also shows that an attention-based BLSTM-RNN can outperform a CNN for end-to-end neural network based automated speech scoring [18]. We propose to use three attention BLSTM-RNNs to learn the mapping functions between the spoken responses and the three dimensions (content, delivery and language use) of scoring rubrics. The schematic diagram of the proposed NN architecture for scoring spoken responses is shown in Figure 1.

**Language Use** Vocabulary, grammar and syntax related features extracted from ASR word hypotheses are generally used to measure the appropriateness of language use. The non-native ASR system used here was trained by using a large database, covering as many accents or L1s as possible, to recognize non-native speech. In [27], it is shown that syntactic, morphological and dependency-related information about words provides useful information for word-level detection of grammatical errors made by machine translation systems. Inspired by this work, we concatenate multiple one-hot vectors representing Part-Of-Speech (POS), such as verb or noun, syntactic dependency labels (DEP) that describe the relations between words, such as subject or object, and morphology (Morph.) features obtained with spaCy[1], a Python-based industrial-strength natural language processing tool. Figure 2 shows an example of input features for the word 'read' to the NN of language use. In total, there are 19, 51 and 248 word-level labels for POS, DEP and Morph., respectively.

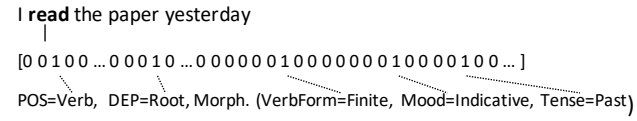


**Fig. 2:** Binary vector for representing the features of POS, DEP, and Morph.

**Content** This aspect of the scoring rubric assesses the topical relevance and content appropriateness. The features used for the content NN are extracted from the word sequence recognized by the same non-native ASR system used for language use. We propose a prompt-aware BLSTM-RNN to content scoring,

$$h_t = \mathcal{H}_{LSTM}(W_{eh}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

$$x_t = \{e_t^r, v^p\} \quad (2)$$

where each word $w_t^r \in \{w_1^r, w_2^r, \dots, w_T^r\}$ or $w_t^p \in \{w_1^p, w_2^p, \dots, w_N^p\}$ in the response or prompt (listening material) is mapped to word embedding, $e_t^r$ or $e_t^p$, via the embedding layer, which is initialized by Google's Word2Vec (300-dimensional vector) and optimized during model training; the word sequence contained in the prompt is encoded into a fixed length prompt-vector, $v^p$, in a low-dimensional space by a BLSTM-RNN; the prompt-vector, $v^p$, is employed as a condition by appending to the word embedding, $e_t^r$, of the response and fed into a BLSTM-RNN layer. The prompt-

[1] https://spacy.io/

aware NN approach only needs to train a generic model for the spoken responses to different prompts instead of prompt-specific models, i.e., the responses to each prompt in the training set are used to train a model. This approach effectively predicts the scores for the responses to the unseen prompts in the test set, i.e. the prompt has no corresponding responses in the training set, since it can learn commonalities across the training responses to different seen prompts and do interpolation for the unseen prompts.

**Delivery** It measures the pronunciation, stress, fluency, and intonation of spoken responses. The features to the inputs of the delivery NN are obtained via a forced alignment process, which uses native ASR model to align the recognition hypotheses generated by non-native ASR to audio recordings to automatically produce different level segmentation. The features include duration, pitch, intensity, following silence or pause length, posterior probabilities of AMs from both native ASR and non-native ASR, LM score from non-native ASR, and confidence score of non-native ASR. An eight-dimensional continuous vector is employed to represent these features averaged by the frame length. In this study, the performance of the averaged features at phoneme, syllable and word levels will be tested.

Generally, the first state, $h_1$, and the last state, $h_T$, of the BLSTM-RNN is averaged as the final prediction. To utilize the information from the contextual states, i.e., time steps, $h_t,\ t\in\{1,2,\ldots,T\}$ in both forward and backward directions, we add a feed-forward attention [25] layer to the outputs of the BLSTM for determining which states to pay attention to for the regression layer. It can produce a single vector $z$ from an entire state sequence as

$$\alpha_t = \frac{\exp\big(\alpha(h_t)\big)}{\sum_{k=1}^{T}\exp\left(\alpha(h_k)\right)} \tag{3}$$

$$z = \sum_{t=1}^{T}\alpha_t\, h_t \tag{4}$$

where vectors $h_t$ in the state sequence are fed into a learnable function $\alpha(h_t)$ to produce a probability vector $\alpha$. The vector $z$ is computed as a weighted average of $h_t$, with weights given by $\alpha$. It is implemented as a merged layer by applying the multiply operation on the outputs of the BLSTM layer and the outputs of the attention layer in Keras[2]. The mean of the merged layer is the predicted score for the response. Three subscores are concatenated together and fed into a dense layer to predict a holistic score to each response.

The dialogue contains multi-turn interactions. Sometimes, it is difficult to understand the current utterance without the context in a conversation. To capture information about sequential utterances from both the prompt and the response, we propose to use a MemN2N, illustrated in Figure 3, where the attention to the contextual prompts and responses can be automatically learned in an end-to-end manner, for the content scoring of spoken dialogue responses. The MemN2N has been successfully applied to many tasks, modeling long-term dependencies in sequential data, such as question answering [19] and spoken dialogue understanding [28]. Equation 2 is modified to

$$x_t = \{e_t^r,\ v^p,\ a^p \cdot v_h^p,\ a^r \cdot v_h^r\} \tag{5}$$

where $a^p$ and $a^r$ are the attention vectors to the history prompts $v_h^p$ and history responses $v_h^r$. The model predicts a content subscore for each response in the dialogue. A single subscore for the entire conversation is obtained by averaging all subscores of the responses. The same strategy is applied to the subscores of delivery and language use.

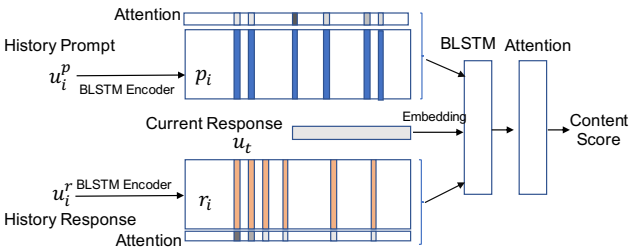


**Fig.3:** *The end-to-end memory networks (MemN2N) for content scoring of spoken dialogue responses*

## 5. EXPERIMENTS

### 5.1. Experimental Setup

The monolog corpus contains 8,738 responses from 3,225 test takers from 14 native language backgrounds (L1s). The dialog corpus consists of 1,847 conversations (10,865 responses) from 1,847 test takers representing 51 L1s. The scores are rated by at least 2 experts. A third or fourth opinion is given when the scores from those two experts are different. The final adjudicated scores are used as the reference scores to build the scoring model. Responses scored with zero are generally non-English responses, off-topic responses, or responses with no intelligible speech. Both corpora are divided into training and test sets (with no speaker overlap) for the current study. The corresponding number of speakers and the number of responses are presented in Table 1.

Table 1: *Number of speakers and number of responses for each data partition*

| Partitions | Monolog | | Dialog | |
|---|---|---|---|---|
| | Speakers | Responses | Speakers | Responses |
| Train | 2,511 | 6,635 | 1,567 | 9,380 |
| Test | 714 | 2,103 | 280 | 1,485 |

Two non-native ASR systems are used to generate recognized hypotheses for monolog and dialog corpora. AMs were trained by iVector-based BLSTM-RNNs with 160 hours and 800 hours of non-native speech recorded by 1,600 children and 8,700 adults worldwide. The LMs are adapted using the task prompts. The details about the construction of these two ASR systems and system performance are presented in [5, 10]. A native AM trained by LibriSpeech [29] (containing approximately 960 hours of speech collected from 2,338 speakers) is employed for the forced-alignment with the recognition hypotheses for generating delivery features. The recognition hypotheses of spontaneous speech contain filler words and repeated partial words. These were removed to generate language use and content features but retained for precise forced-alignment to produce delivery features.

The baseline system is built with the features extracted using SpeechRater [1], a conventional automated speech scoring engine. There are over 100 features covering fluency, rhythm, intonation, stress, pronunciation, grammar, vocabulary use, content and etc., which are mostly aggregated over a whole response, e.g., speaking rate, number of chunks, global AM and LM scores (normalized), Content Vector Analysis [6]. A detailed description of these features is provided in [1,30]. The different regression methods provided by the SKLL[3] toolkit, including Logistic Regression, AdaBoost, Decision Tree, Gradient Boost, Support Vector Machine, Random

[2] https://keras.io/
[3] https://github.com/EducationalTestingService/skll

Forest, etc., were employed to build scoring models. The hyper-parameters of these regressors were optimized by SKLL internally via cross-validation on the training set. Among all these regressors, the Random Forest Regressor achieves the highest performance. We hereafter use the predictions from Random Forest Regressor as the output of the baseline system.

The neural approaches and the features introduced in Section 4 are used to build the neural-based scoring systems. The Adam optimization algorithm is used to update the network parameters towards minimizing the loss function of mean squared error (MSE) between the predicted scores and reference scores over the training set. We shuffle the training samples and select 20% of them as a development set. Instead of using early stopping methods, we train the model for a fixed number of epochs. We set model checkpoints, i.e., we save the model weights after each epoch if the performance of the model on the development set is improved, and store them in a callback list during training. We select the model with the best performance in the callback list as the final model. To avoid overfitting, we employ dropout (with p=0.5). The number of units for BLSTM is 128; A batch size of 64 samples is used in each epoch; 100 epochs are used for model training; The memory size of MemN2N is 20 to store carried information from previous 10 turns of prompts and responses.

### 5.2. Results and Discussion

The performance of automated speech scoring system is evaluated using Pearson correlations between the predicted scores and the reference scores. The predicted scores produced by the systems are continuously valued scores while the human experts rate the spoken responses using scoring rubrics on a discrete point scale. We round the scores generated by the systems to the nearest integers and then calculate their correlations with human scores.

Table 2: *Correlations of automatically predicted subscores by different NN-based models with reference scores across different tasks*

| | Delivery | | | Lang | Content | |
|---|---|---|---|---|---|---|
| | Phn | Syl | Word | use | w/o $v^p$ | w/ $v^p$ |
| Mono | 0.531 | 0.565 | 0.553 | 0.656 | 0.709 | 0.723 |
| Dial | 0.570 | 0.591 | 0.579 | 0.676 | 0.713 | 0.741(0.749) |

Table 3: *Correlations of the holistic scores predicted by baseline systems and neural-based systems with reference scores across different tasks*

| | Baseline | Neural | Human-Human |
|---|---|---|---|
| Mono | 0.684 | 0.747 | 0.715 |
| Dial | 0.691 | 0.772 | 0.702 |

We observed that adding an attention layer after the LSTM layer brings a substantial performance improvement, i.e., the increment in the correlation coefficients ranges from 0.01 to 0.05 across different NN-based subscore models of monologic and dialogic tasks. Due to the limited space, the detailed results of the comparison for with or without attention layer are not listed here. Table 2 shows the performances of the different attention BLSTM-RNNs for modeling subscores, *Delivery, Language use, and Content*, with different-level inputs, *Phone, Syllable and Word*, or different conditions, *without or with prompt-encoder (MemN2N encoder)*, across different tasks, *monolog and dialog*. The performance of predicted subscores is also measured by their correlations with the holistic reference scores since there are no responses with manually marked subscores in both corpora.

**Delivery** The syllable-level features outperform the phone-level features and word-level features for predicting delivery subscore for the testing sets of both monolog and dialog corpora. Here the delivery feature like LM score for syllable or phone is employed the LM score of the word which syllable or phone belongs to. The syllable, which is typically made up of a vowel with the optional proceeding and succeeding consonants and has relatively invariant duration, can be used to calculate speaking rate more precisely over the word. The syllable is also the unit that carries stress, which influences the rhythm or prosody to help deliver the information to the audience.

**Language Use** The subscores predicted from the attention BLSTM-RNN with word-level POS, DEP and Morph in a binary vector representation have correlations of 0.656 and 0. 676 with the reference scores for the responses of monolog and dialog, separately. Word embeddings are also useful to capture the lexical and syntactic information. Here our assumption is that POS, DEP and Morph labels can strip out the semantic (or content) information contained in the word embeddings and is thus more focused on language use like grammar.

**Content** Using the vectors produced by prompt-encoder as conditional inputs (by appending to word embeddings) to attention LSTM-RNN can improve the performance of content grading, i.e., the correlations are improved from 0.709 to 0.723 for monolog and from 0.713 to 0.741 for dialog. The MemN2N can further improve the correlation coefficient to 0.749 for dialog. We find that the prompt information contributes more in the dialog than in the monolog to distinguish between the high-scoring samples and low-scoring samples. In addition, we observed that word embeddings refined in the training of scoring model can slightly outperform the fixed embeddings.

Table 3 presents the correlations of the holistic scores predicted by baseline systems and neural-based systems with reference scores. The correlation improvements achieved by our proposed neural approaches are 0.063 and 0.081 over the baseline for monolog and dialog, respectively. The improvements are significant, i.e., the z-scores of Steiger's z-test [31] achieve 4.26 and 2.17. The correlations between the automated scores and expert scores are now superior to the human-human agreement levels for both tasks. The results shown in Table 3 are obtained by combining three subscore NNs at the last output layer. We also tried merging them at the first layer or the middle layer. The preliminary results show that the fusion of input features can slightly outperform that of final outputs in terms of generating holistic score since the relations of the input features across different dimensions can also be learned by NN. However, it cannot address the appropriateness of the different aspects of spoken language proficiency, represented by subscores.

## 6. CONCLUSIONS

In this paper, we have proposed using various neural approaches: attention, BLSTM-RNN, encoder, MemN2N, and conditional inputs, for automated grading of spoken monolog and dialog responses. Our approaches can capture the evolution (or trajectory) of the input features, such as sequential word embeddings and sequential acoustic vectors. They show a superior scoring performance compared to the traditional approaches, which are limited by conventional machine learning methods and only can use handcrafted aggregated features over the whole response or have to normalize the variable-length feature sequence to a fixed length. In the future, we will improve the interpretability of neural networks for revealing what these scoring models are really learning and thus investigate the potentials to provide diagnostic targeted feedback, e.g., which word is mispronounced or has a grammatical error.

## 7. REFERENCES


[1] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, pp. 883-895, 2009.

[2] D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech & Language*, vol. 25, no. 2, pp. 282–306, 2011.

[3] S. Xie, K. Evanini and K. Zechner, "Exploring content features for automated speech scoring," in *Proc. of NAACL HLT*, 2012.

[4] A. Metallinou and J. Cheng, "Using Deep Neural Networks to improve proficiency assessment for children English language learners," in *Proc. of Interspeech,* pp. *1468–1472,* 2014.

[5] Y. Qian, X. Wang, K. Evanini and D. Suendermann-Oeft , "Self-adaptive DNN for improving spoken language proficiency assessment," in *Proc. of Interspeech*, pp. 3122-3126, 2016.

[6] K. Evanini, S. Singh, A. Loukina, X. Wang and C. M. Lee, "Content-based automated assessment of non-native spoken language proficiency in a simulated conversation," in *NIPS Workshop on Machine Learning for Spoken Language Understanding and Interaction*, 2015.

[7] R. van Dalen, K. Knill, and M. Gales, "Automatically grading learners' English using a Gaussian process," in *Proc. of SLaTE*, 2015.

[8] Y. Wang, M. Gales, K. Knill, K. Kyriakopoulos, A. Malinin, R. van Dalen and M. Rashid., "Towards automatic assessment of spontaneous spoken English," *Speech Communication*, Vol. 104, pp. 47-56, 2018.

[9] W. Xiong, K. Evanini, K. Zechner and L. Chen, "Automated content scoring of spoken responses containing multiple parts with factual information," in *Proc. of the Workshop on Speech and Language Technology in Education*," pp. 137-142, 2013.

[10] Y. Qian, K. Evanini, X. Wang, C. M. Lee and M. Mulholland "Bidirectional LSTM-RNN for improving automated assessment of non-native children's speech," in *Proc. of Interspeech,* pp. 1417–1421, 2017.

[11] F. Dong and Y. Zhang, "Automatic features for essay scoring – an empirical study," In *Proc. of EMNLP*, pp. 1072–1077, 2016.

[12] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," In *Proc. of EMNLP*, pp.1882–1891, 2016.

[13] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," in *Proc. of ACL*, vol. 1, pp. 715–725, 2016.

[14] B. Riordan, A. Horbach, A. Cahill, T. Zesch and C.M. Lee, "Investigating neural architectures for short answer scoring" in *Proc. of the 12th Workshop on Innovative Use of NLP for Building Educational Applications,* pp. 159-168, 2017.

[15] C. M. Lee, S.-Y. Yoon, X. Wang, M Mulholland, I. Choi and K. Evanini, "Off-topic spoken response detection using Siamese Convolutional Neural Networks," in *Proc. of Interspeech*, 2017.

[16] A. Malinin, K. Knill, A. Ragni, Y. Wang and M. Gales., "An attention based model for off-topic spontaneous spoken response detection: an Initial Study," in *Proc. of SLaTE,* 2017.

[17] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *Proc. of ASRU*, pp. 338–345, 2015.

[18] L. Chen, J. Tao, S. Ghaffarzadegan and Y. Qian, "End-to-end neural network based automated speech scoring", in *Proc. of ICASSP*, 2018.

[19] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, "End-to-end memory networks," in *Proc. of NIPS*, pp. 2431–2439, 2015.

[20] K. Evanini, and X. Wang, "Automated speech scoring for non-native middle school students with multiple task types", in *Proc. of Interspeech*, pp. 2435-2439, 2013.

[21] J. L. Elman, "Finding structure in time," *Cognitive Science*, 14(2):179–211, 1990.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9 (8), pp. 1735-1780, 1997.

[23] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. of ICASSP*, pp. 4960–4964, 2016.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate,"arXiv:1409.0473, 2014.

[25] C. Raffel and D. P. W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," *arXiv:1512.08756*, 2015.

[26] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. of EMNLP*, 2014.

[27] T. Arda, H. Véronique and M. Lieve, "A neural network architecture for detecting grammatical errors in statistical machine translation," *Prague Bulletin of Mathematical Linguistics*, vol. 108, no. 1, pp. 133-145, 2017.

[28] Y.-N. Chen, D. Hakkani-Tur, G. Tur, J. Gao, and L. Deng, "End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding," in *Proc. of Interspeech*, pp. 3245–3249, 2016.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. of IEEE ICASSP*, pp. 5206–5210, 2015.

[30] L. Chen, et al, "Automated scoring of nonnative speech using the SpeechRater$^{SM}$ v. 5.0 Engine", ETS Research Report Series, 2018.

[31] I. A. Lee and K. J. Preacher. Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. 2013. Available from http://quantpsy.org.