INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Bidirectional LSTM-RNN for Improving Automated Assessment Of Non-native Children's Speech

Yao Qian, Keelan Evanini, Xinhao Wang, Chong Min Lee, Matthew Mulholland
Educational Testing Service Research, USA
{yqian,kevanini,xwang002}@ets.org

Abstract

Recent advances in ASR and spoken language processing have led to improved systems for automated assessment for spoken language. However, it is still challenging for automated scoring systems to achieve high performance in terms of the agreement with human experts when applied to non-native children's spontaneous speech. The subpar performance is mainly caused by the relatively low recognition rate on non-native children's speech. In this paper, we investigate different neural network architectures for improving non-native children's speech recognition and the impact of the features extracted from the corresponding ASR output on the automated assessment of speaking proficiency. Experimental results show that bidirectional LSTM-RNN can outperform feed-forward DNN in ASR, with an overall relative WER reduction of 13.4%. The improved speech recognition can then boost the language proficiency assessment performance. Correlations between the rounded automated scores and expert scores range from 0.66 to 0.70 for the three speaking tasks studied, similar to the human-human agreement levels for these tasks.

Index Terms: non-native children's speech, speech recognition, bidirectional LSTM-RNN, DNN and i-vector

1. Introduction

Automated assessment of several aspects of spoken language proficiency, including vocabulary, grammar, content appropriateness, and discourse coherence, depends heavily on how accurately the input speech can be recognized. Limited by low recognition accuracy, especially for children's speech, early automated language assessment systems for children's speech mainly focus on reading assessment, for example, the Reading Tutor from CMU's project LISTEN [1], IBM's Reading Companion [2], among others [3,4]. Generally, the users are asked to engage in a read-aloud task and the systems provide feedback to the users based on the overall accuracy of their reading, the metrics associated with pronunciation, fluency and prosody, or the specific types of reading errors that the user made.

Automated systems for assessing responses to a wider variety of speaking tasks, such as picture narration and source-based open questions, have appeared recently, and these systems can provide a more comprehensive evaluation of the speakers' communicative competence. For example, [5] investigated the performance of an automated speech scoring system applied to the TOEFL Junior Comprehensive assessment, which was designed to evaluate English communication skills of students aged 11 and older. In addition, [6] investigated automated speech scoring for the AZELLA speaking test, which contains a variety of spoken tasks used for assessing the English speaking proficiency of K-12 students. The state-of-the-art acoustic modeling based on deep neural networks (DNN) have significantly improved speech recognition performance. The improved speech recognition performance in turn improves the quality of the extracted features for automatic assessment of spoken language proficiency [6-9]. It has largely boosted the applications of language proficiency assessment in terms of human-machine spoken communication.

However, the above assessment systems struggle to achieve high performance (in terms of agreement with human experts) when the input is non-native children's spontaneous speech. Children have shorter vocal tracts and smaller vocal folds, which lead to higher fundamental frequencies and formant frequencies than for adults [10-12]. Children's speaking rates tend to be slower and more variable overall due to the fact that their articulators have not fully developed yet [13]. These facts cause difficulties for the automated scoring of pronunciation, fluency and prosody. In addition, children's choices of vocabulary and syntax tend to differ from adult patterns, and children may use greater amounts of invented words and ungrammatical phrases [14]. The proficiency assessment systems have to face up to these issues if evaluating on language use and meaning delivery.

Recently there are several studies on using deep learning technologies to improve acoustic modeling of children's speech [6,15-17]. In [15], a comparison is made between GMM-HMM and DNN-HMM systems using various amounts of training data for recognizing non-native children's speech for spoken language assessment applications. This study found that the DNN models outperform the GMM models when enough training data was available to train the DNN parameters reliably, and that the observed improvement in recognition performance increases monotonically with more training data. In [16], the best results for children's speech recognition were obtained by training on a large amount of data, which better matched children's speech, aided by convolutional, Long Short Term Memory (LSTM), DNN (CLDNN). Many acoustic modeling techniques for improving children's speech recognition shown in the literature, e.g., spectral smooth, VTLN and using pitch features, are found not to be effective, given their voice search task with trained acoustic model by 1.9 million utterances [16]. The use of DNN-HMM is also explored to improve speech recognition performance for a better assessment of children English language learner [6]. However, to our knowledge, there has been little research investigating deep neural network architectures for the purpose of improving automatic proficiency assessment of non-native children's speech.

Recurrent Neural Networks (RNN), especially with bidirectional LSTM cells [18-22], in principle can capture information from anywhere in the feature sequence. It is attractive for some applications in which the real-time

http://dx.doi.org/10.21437/Interspeech.2017-250

requirement is low. In this paper, we explore the gain of using bidirectional LSTM-RNN in contrast to feed-forward DNN for automated scoring of children's spoken responses.

2. Data and Task

In this study, we use a corpus that contains non-native children's speech drawn from a pilot version of the TOEFL Junior Comprehensive assessment administered in late-2011 [5]. The TOEFL Junior Comprehensive is a computer-based test containing four sections: Reading Comprehension, Listening Comprehension, Speaking, and Writing. It is intended for middle school students around the ages of 11-15, and is designed to assess a student's English communication skills through a variety of tasks. This study focuses on the Speaking section, which contains the following three task types eliciting spoken responses:

- Read Aloud (RA): the test taker reads a paragraph (containing approximately 90 - 100 words) presented on the screen out loud
- Picture Narration (PN): the test taker is shown six images that depict a sequence of events and is asked to narrate the story in the pictures
- Listen Speak (LS): the test taker listens to an audio stimulus (approximately 2 minutes in duration) containing information about a non-academic topic (for example, a homework assignment) or an academic topic (for example, the life cycle of frogs) and provides a spoken response containing information about specific facts in the stimulus

Each speaker provided 5 responses: one RA, one PN, and three LS. The responses to each of the three task types are approximately 60 seconds in duration. The corpus includes responses from 3,385 test takers from the following native language backgrounds: Arabic, Chinese, French, German, Indonesian, Japanese, Javanese, Korean, Madurese, Polish, Portuguese, Spanish, Thai, and Vietnamese. It is divided into the following five sets (with no speaker overlap) for the current study: ASR training (AsrTrain), ASR development (AsrDev), ASR evaluation (AsrEval), Scoring Model training (SMTrain) and Scoring Model evaluation (SMEval). The corresponding number of speakers, number of responses, and hours of speech are presented in Table 1. Non-scorable responses (e.g., responses with large amounts of background noise) and responses scored with zero (e.g., non-English responses and off-topic responses) are excluded in from this study. Each response is scored on a scale of 1-4 by at least 2 expert human raters following scoring rubrics on fluency and reading accuracy for the RA task, and content, delivery and language use for PN and LS tasks [23].

Table 1: *Number of speakers, number of responses, and duration of speech for each data partition in the children's corpus*

Partition	Speakers	Responses	Duration (hrs)
AsrTrain	1,625	7,594	137.2
AsrDev	30	150	2.5
AsrEval	30	150	2.5
SMTrain	967	4,334	81.7
SMEval	733	3,302	62.0

3. Speech Recognizers for Children

We built speech recognizers based on the structures: feed-forward DNN and bidirectional LSTM-RNN (BLSTM-RNN) for children's speech.

A GMM-HMM is first trained to obtain senones (tied tri-phone states) and the corresponding aligned frames for DNN training. The input feature vectors used to train the GMM-HMM contain 13-dimensional MFCCs and their first and second derivatives. Contextually dependent phones, tri-phones, are modeled by 3-state HMMs and the pdf of each state is represented by a mixture of 8 Gaussian components. The splices of 9 frames (4 on each side of the current frame) are projected down to 40-dimensional vectors by linear discriminant analysis (LDA), together with maximum likelihood linear transform (MLLT), and then used to train the GMM-HMM using ML. To alleviate the mismatch between the training criterion and performance metrics, the parameters of the GMM-HMM are then refined by discriminative maximum mutual information (MMI) training.

An i-vector is a popular auxiliary feature for improving DNN-based ASR. It is a compact representation of a speech utterance that encapsulates speaker characteristics in a low-dimensional subspace [24, 25] and has become the state-of-the-art approach in the field of speaker recognition. Using i-vectors is also a promising approach to speaker adaptation for speech recognition and appending the i-vector to frame-level acoustic features has been reported to improve the performance of ASR based on DNN acoustic modeling [26-28]. The AsrTrain partition of the children's corpus is also used to train the following hyper-parameters: GMM-UBM and T-matrix for i-vector extraction.

The features used to train the DNN are concatenated MFCC features and i-vector features. The MFCC features have the same dimensions as those used in GMM-HMM, while the i-vector features have 100 dimensions extracted from each response and are appended to the frame-level MFCC features. The input features stacked over a 15 frame window (7 frames to either side of the center frame for which the prediction is made) are used as the input layer of DNN. The output layer of the DNN has 3,957 nodes, i.e., the senones of the HMM obtained by decision-tree based clustering. The input and output feature pairs are obtained by frame alignment for senones with the GMM-HMM. The DNN has 5 hidden layers, and each layer contains 1,024 nodes. The sigmoid activation function is used for all hidden layers. All the parameters of the DNN are first initialized by layer-wise back-propagation pre-training, then trained by optimizing the cross-entropy function through back-propagation, and finally refined by sequence-discriminative training, state-level minimum Bayes risk (sMBR).

The input features to the BLSTM-RNN are the same as those to DNN, i.e., 40-dim MFCC and 100-dim i-vector, but there is no stacked frame window since the RNN architecture already captures the long-term temporal dependencies between the sequential events [29]. A three layer stacked bi-directional LSTM is employed and uses recurrence connections with delays -1 for the forward and 1 for the backward at the first layer, -2 for the forward and 2 for the backward at the second layer, and -3 for the forward and 3 for the backward at the third layer. We employ the architecture in which each LSTM layer has a linear recurrent projection layer [22]. 40 left and right contexts are set as the number of steps used in the estimation of LSTM state. The number of LSTM cell is 640 and the number of recurrent projection units is 128. A back-propagation through time (BPTT) learning algorithm is used to train BLSTM-RNN parameters. The number of time steps to back-propagation is 20 (T_{bptt}). For deep BLSTM, BPTT algorithm is applied to both

forward and backward hidden nodes, and back-propagates layer by layer. The output states inventory is 3,957, which is the same as that of DNN. No delay (zero frame) is used to predict for the LSTM output state label. The BLSTM-RNN parameters are also refined by sMBR training.

The CMU pronunciation dictionary [30] is used to build a grapheme-to-phoneme (G2P) converter by data-driven joint-sequence models [31]. After text normalization for the transcriptions, we use G2P to automatically generate pronunciations for the words in the transcription not contained in the CMU dictionary and combine them with the CMU dictionary to create a new pronunciation dictionary. A trigram LM is trained from the transcriptions of the AsrTrain set using the SRILM toolkit [32]. LM is represented as a finite state transducer (FST) for weighted FSTs based decoding.

Speech recognizers for children's speech were constructed using the Kaldi toolkit [33]. Although there are many acoustic and linguistic mismatches between the speech of adults and children, our previous work [17] shows that adult speech can still be used to boost the performance of a speech recognizer for children using acoustic modeling techniques based on the DNN framework, which can represent high-level abstractions of complex data set and learn the commonalities between adults and children's speech. A corpus of non-native adult speech drawn from the TOEFL iBT assessment is thus used to enhance the speech recognition performance. We build four ASR systems as follows:

- DNN: Feed-forward DNN based speech recognizer with i-vectors. The training data used for this system is the AsrTrain partition of the children's corpus.
- RNN: BLSTM-RNN based speech recognizer with i-vectors. The same training data as that of DNN is used.
- DNN_C: Feed-forward DNN based speech recognizer with i-vectors. The training data is a combination of the AsrTrain partition of the children's corpus and 10,000 responses (~150 hours) randomly selected from the adult corpus. It shows that combining approximately the same size of adult data can achieve the best recognition performance [17].
- RNN_C: BLSTM-RNN based speech recognizer with i-vectors. The same training data as that of DNN_C is used.

The performance in terms of word error rate (WER) is reported on the AsrEval set of the children's corpus. Table 2 shows the WERs obtained by the four ASR systems (DNN, RNN, DNN_C and RNN_C) across the three tasks (RA, PN and LS). The ASR system built by the acoustic model trained by BLSTM-RNN achieves significantly better performance than that of DNN with an overall WER reduction from 22.4% to 19.4%. The ASR performance for the three speaking tasks studied are all improved. Among those three tasks, the LS task shows the largest gain with a WER reduction from 27.6% to 23.7%. The frame accuracy is often used to evaluate the performance of a DNN by isolating the issues caused by the vocabulary and the language model in the ASR system. Figure 1 shows the frame accuracy across the training iterations obtained by DNN and RNN based systems on the AsrDev set of the children's corpus. BLSTM-RNN can improve the frame accuracy from 57.5% to 66.7% at the final iteration over the feed-forward DNN. It indicates that BLSTM-RNN can produce sharper frame posteriors than those of DNN. In the next section, we will investigate whether this accuracy gain can translate into a gain of automatic assessment performance.

Although RNN_C achieves the best performance among these four systems, RNN_C is only slightly better than RNN. More specifically, WER is reduced from 19.4% to 19.1%, which indicates that adding adult speech into the training set is marginally effective in improving the performance of speech recognition for children. However, a different phenomenon is observed for DNN based systems, where DNN_C significantly outperforms DNN, i.e., a relative WER reduction of 10.7% is obtained by combining child training data with adult data. We conjecture that the long-range contexts are captured by BLSTM-RNN and the corresponding language model information carried by adult data is also learned by RNN, which might be mismatched to children's data. The frame accuracy also indicates that the gain from RNN to RNN_C is marginal on the AsrDev set of the children's corpus, i.e., RNN_C achieves the frame accuracy of 67.3% in contrast to 66.7% obtained by RNN.

Table 2: *WERs (%) of four ASR systems across the three tasks*

Systems	RA	PN	LS	Overall
DNN	7.7	24.4	27.6	22.4
RNN	7.1	20.9	23.7	19.4
DNN_C	7.2	21.0	24.8	20.0
RNN_C	7.4	21.0	23.2	19.1

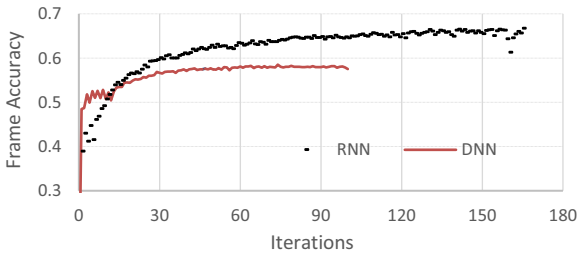

Figure 1. *Frame accuracy (%) of DNN and RNN systems on the AsrDev set of children corpus across the training iterations*

4. Language Proficiency Assessment

Over 100 features covering a range of linguistic characteristics related to English speaking proficiency were extracted using the SpeechRater system [34]. The feature extraction is performed by using a two-pass approach that first conducts ASR on the spoken response using acoustic and language models trained from non-native spontaneous speech and then conducts forced alignment of the spoken response to the ASR output using an acoustic model trained on native speech. The non-native ASR is mainly used to extract the features that address the appropriateness of content, vocabulary, grammar and context usage, while the forced alignment based on the native ASR is employed to extract the features for evaluating delivery (pronunciation, fluency, and intonation). The scoring model is built on the SMTrain set by using a random forest (RF) regressor and evaluated on the SMEval set. The scores for each task are mostly rated by two experts except that a third or fourth opinion is given when the scores from those two experts are different. We use the final adjudicated scores as the reference score to build the scoring model. Reference score distribution for the responses of all task types is listed in Table 3. The performance of automatic assessment system is evaluated by the Pearson correlations between these predicted sores and the reference scores.

Table 3: *Human score distribution for the responses of all tasks*

Score	1	2	3	4
Percentage (%)	19.2	41.6	28.9	10.3

In this study, we explore the impact of features extracted from different neural network architectures on the automated assessment of language proficiency. Features assessing the following five dimensions of speaking proficiency are included in the automated scoring models: Content, Grammar/ Vocabulary, Pronunciation, Fluency and Prosody. Features extracted from non-native ASR systems are investigated through a comparison of the DNN- and RNN-based results; features extracted using the ASR system trained on native speech will be investigated in a subsequent study. The features that resulted in the highest correlations with reference scores across these five dimensions are as follows; the corresponding correlation coefficients with the reference scores are shown in Table 4.

Content: In the context of the RA task, content quality is based on the number of correctly read words per minute, a metric that is strongly correlated with comprehension [35]. Content vector analysis (CVA) is employed for the PN and LS tasks. A test taker's spoken response is graded by the cosine similarity between recognized word sequence from the test taker's response and the training set grouped by the human scores. Word sequence is represented by a vector.

Grammar/Vocabulary: Log probability from LM.

Pronunciation: Normalized posterior probability from AM

Fluency: Number of pauses per word

Prosody: Relative frequency of stressed syllables in percentage

Table 4: *Correlations of the highest performing features from five dimensions across three speaking tasks for DNN- and RNN-based ASR systems*

	RA		PN		LS	
	DNN	RNN	DNN	RNN	DNN	RNN
Content	0.65	0.68	0.39	0.40	0.41	0.42
Grammar	0.52	0.50	0.36	0.36	0.49	0.45
Pronunciation	0.49	0.51	0.50	0.53	0.57	0.60
Fluency	0.50	0.52	0.48	0.52	0.46	0.49
Prosody	0.45	0.47	0.46	0.50	0.47	0.49

Table 4 indicates that the highest performing features generated from RNN-based ASR typically outperform those from DNN-based ASR, except for the log probability of LM for PN and LS tasks. The same phenomenon is observed for several other features extracted by using LM scores. We think that the RNN architecture can capture the long-range temporal dynamics and result in the relative weak LM needed for ASR since AM already carries certain contextual information. The quality of the features, which are highly related to the frame posteriors, is largely improved by using RNN-based ASR in comparison with DNN-based ASR. The sharper frame posteriors can improve the syllable and word boundaries detection and thus make the calculation for the number of pauses and the recognition of stressed syllables more accurate.

All features generated by DNN and RNN based non-native ASRs and native ASR are fed into the training of scoring models and two automated assessment systems, DNN-based and RNN-based, are built. The correlations of predicted scores by different automated assessment systems with the reference scores across different tasks are shown in Table 5. The RNN-based system outperforms the DNN-based system across all tasks studied. The largest gain is obtained by PN task among the three tasks, i.e., the correlation is improved from 0.71 to 0.73. The predicted scores produced by the systems are continuously valued scores while the experts rate the spoken responses using scoring rubrics on a discrete 4-point scale. Table 5 also presents the correlations between human scores and predicted system scores rounded to the nearest integer in brackets. The automated scores from RNN-based system have a correlation of 0.70 with reference scores, which is similar to human-human agreement level (r = 0.71) for the LS task. Although the correlation improvement achieved by the RNN system is 0.02 over the DNN system, it is still a substantial improvement compared to the gap (0.03) of correlation between human-machine and human-human, i.e., the z-score of Steiger's Z test achieves 2.3 [36].

By analyzing the confusion matrices of predicted scores and reference scores, we find the most confusable scores are between 3 and 4, and the recall of score 4 is pretty low. An example of confusion matrix on the LS task of SMEval set is shown in Table 6. We think it is most likely caused by the score distribution, as shown in Table 3, where the distribution of score 4 is the lowest, used for training scoring models. The most appropriate algorithm for automated scoring can be further investigated in the future.

Table 5: *Correlations of automatically predicted scores (rounded scores) by the features generated by different non-native ASRs, with reference scores across different tasks*

	DNN	RNN	Human-human
RA	0.73 (0.66)	0.74 (0.67)	0.71
PN	0.71 (0.64)	0.73 (0.66)	0.70
LS	0.75 (0.68)	0.76 (0.70)	0.71

Table 6: *Confusion matrix of the predicted scores by RNN based system on the LS task of SMEval set (rows: references; column: predictions)*

	1	2	3	4	Total
1	175	235	10	0	420
2	47	599	153	6	805
3	1	159	332	26	518
4	0	11	114	45	170
Total	223	1004	609	77	1913

5. Conclusions

The performance of the features used for scoring models in automated assessment of language proficiency depends largely upon how accurately the test taker's input speech can be recognized. The bidirectional LSTM-RNN can significantly improve the recognition performance of non-native children's speech, which in turn improves automated scoring performance for all three speaking tasks studied. Among the three speaking tasks, the listen speak task shows the highest correlation of 0.70 between automated scoring and expert scoring, which is pretty close to the correlation of 0.71 between two human experts. The sharper posteriors produced by RNN-based ASR largely boost the quality of the associated features used for automated scoring models. The more sophisticated features, e.g., coherence addressing the communicative competence, and the features extracted from native ASR, will be investigated to further enhance the performance of automated assessment of spoken language proficiency in the future.

6. References

[1] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: The role of multiple information sources," in *Proc. of the IEEE International Workshop on Multimedia Signal Processing*, 2007.

[2] A. Kantor, M. Cernak, J. Havelka, S. Huber, J. Kleindienst, and D. B. Gonzalez, "Reading Companion: The technical and social design of an automated reading tutor," in *Proc. of the INTERSPEECH Workshop on Child, Computer, and Interaction*, 2012.

[3] K. Zechner, J. Sabatini, and L. Chen, "Automatic Scoring of Children's Read-Aloud Text Passages and Word Lists", *in Proc. of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 10-18, 2009.

[4] R. Downey, D. Rubin, J. Cheng, and J. Bernstein, "Performance of Automated Scoring for Children's Oral Reading", in *Proc. of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 46-55, 2011.

[5] 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.

[6] 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.

[7] Y. Qian, X. Wang, K. Evanini, and D. Suendermann-Oeft, "Self-adaptive DNN for Improving Spoken Language Proficiency Assessment," in *Proc. of INTERSPEECH*, 2016.

[8] 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, 2015.

[9] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *proc. of IEEE ICASSP*, 2015.

[10] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance", in *Proc. of IEEE ICASSP*, 1998.

[11] L. Måhl, "Speech recognition and adaptation experiments on children's speech", Master of Science thesis at the Department of Speech, Music and Hearing, KTH (The Royal Institute of Technology).

[12] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in *Proc. of the INTERSPEECH Workshop on Child, Computer, and Interaction*, 2014.

[13] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. of EUROSPEECh*, 1997.

[14] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for US English: Child interaction with living-room-electronic-devices," in *Proc. of the INTERSPEECH Workshop on Child, Computer, and Interaction*, 2014.

[15] J. Cheng, X. Chen, and A. Metallinou, "Deep neural network acoustic models for spoken assessment applications," *Speech Communication*, vol. 73, pp. 14–27, 2015.

[16] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proc. of INTERSPEECH*, pp. 1611-1615, 2015.

[17] Y. Qian, X. Wang, K. Evanini, and D. Suendermann-Oeft, "Improving DNN-based Automatic Recognition of Non-native Children's Speech with Adult Speech," in *Proc. of the INTERSPEECH Workshop on Child, Computer, and Interaction*, 2016.

[18] A. Graves, N. Jaitly, and A.-R Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM." In *Proc. of IEEE ASRU*, pp.273-278, 2013.

[19] H. Sepp, and S. Jürgen, "Long short-term memory," *Neural computation*, vol.9, no.8, pp. 1735-1780, 1997.

[20] A. Gers, N. Schraudolph, and S. Jürgen. "Learning precise timing with LSTM recurrent networks," *The Journal of Machine Learning Research*, vol.3, pp. 115-143, 2003.

[21] S. Mike, and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol.45, no.11, pp.2673-2681, 1997.

[22] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. of INTERSPEECH,* pp.338-342, 2014.

[23] https://www.ets.org/s/toefl_junior/pdf/toefl_junior_comprehensive_speaking_scoring_guides.pdf

[24] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front end factor analysis for speaker verification, " *IEEE Trans. Acoustic, Speech, Signal Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[25] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification, " *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp.980 – 988, 2008.

[26] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors", in *Proc. of IEEE ASRU*, 2013.

[27] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription", in *Proc. of IEEE ICASSP*, pp. 6334-6338, 2014.

[28] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 11, 2015

[29] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," Feb. 2014. [Online]. Available: http://arxiv.org/abs/1402.1128

[30] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

[31] M. Bisani, and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, Issue 5, pp. 434-451, 2008.

[32] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", in *Proc. of Intl. Conf. Spoken Language Processing*, 2002.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J Silovsky, G. Stemmer, and K. Vesel, "The kaldi speech recognition toolkit," *in Proc. of IEEE ASRU*, 2011.

[34] 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.

[35] J. Hasbrouck, and G. A. Tindal, Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers. *The Reading Teacher*, 59: 636–644. doi:10.1598/RT.59.7.3, 2006.

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