END-TO-END NEURAL NETWORK BASED AUTOMATED SPEECH SCORING

Lei Chen, Jidong Tao‡, Shabnam Ghaffarzadegan⁎, Yao Qian†
Liulishuo Silicon Valley AI Lab, ‡ Midea America Corporation, ⁎ Robert Bosch LLC, † Educational Testing Service (ETS)

Introduction

- Handcrafted scoring features have been widely used in automated speech assessment
- The challenges related to using handcrafted features
  - no guarantee to obtain optimal features
  - substantial development efforts
- Recent successes of end-to-end deep learning (DL) approaches on various tasks in computer vision, speech and language technology provide a promising direction

Previous research

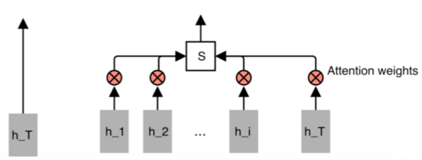
- DL-based ASR improved the automated speech scoring performance by providing more accurate ASR hypotheses and acoustic model (AM) scores [1, 2]
- Increasing number of studies of using Neural Network (NN) methods on rating essays, e.g., [3, 4]
- Limited work on the end-to-end automated speech scoring. For example, [5] tested the learned and the handcrafted features together, while it is not clear if the learned features have independent contributions.

Methods

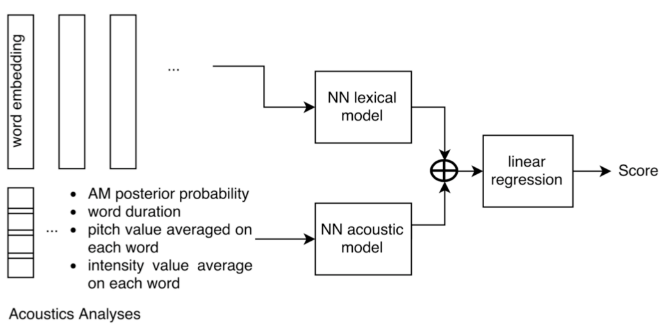
- End-to-end architecture
  - Two DL-based models were used to encode both lexical and acoustic cues

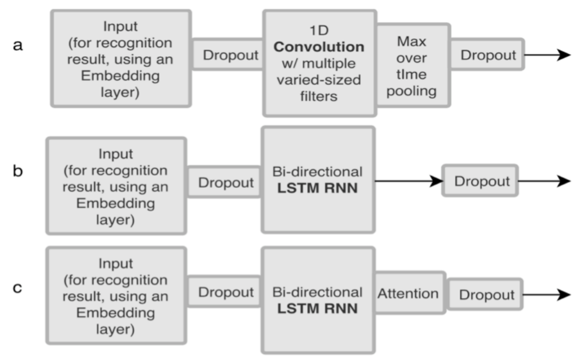
    1) Lexical input: Recognized words were converted to tensors via pre-trained word embeddings

    2) Audio input: On each recognized word, used AM score, word duration, the mean value of pitch, and the mean value of intensity
  - The encoded features were concatenated and fed into a linear regression model to predict scores
- Three types of NN encoders
  - 1D Convolutional Neural Network (CNN)
  - Bi-directional Recurrent Neural Network (RNN) using Long Short-Time Memory (LSTM) cells (BD-LSTM RNN)
  - BD-LSTM RNN using attention mechanism [6]



Experiments





- Data sets
  - TOEFL Practice Online (TPO): English proficiency test used to prepare for the TOEFL test
  - Elicits spontaneous spoken repsonses; 45 - 60 seconds
  - Data partitions: train (2,930), dev (731), eval (1,827)
  - All spoken responses were scored by experienced human raters on a 4-point scale
- ASR
  - DNN-HMM hybrid ASR system based on Kaldi
  - A 5-layer feed-forward DNN AM using features from the current frame plus the previous and following 5 frames
  - Trained on 819 hours of non-naïve spontaneous speech data
- Conventional model
  - Features were extracted by an automated speech scoring system, including fluency, rhythm, intonation & stress, pronunciation, grammar, and vocabulary use
  - Used SKLL toolkit to run machine learning tasks; Gradient Boosting Tree (GBT) model was found to perform best
- NN models
  - were developed by using Keras Python API with Theano as backend
  - 300-dimensional GloVE word embedding vectors
  - Tree Parzen Estimation (TPE) in the Hyperopt Python package for NN hyperparamters tuning
- Evaluation metric
  - Pearson correlation between predicted scores and human rated scores

- CNN model's performance is very close to the conventional model's performance
- BD-LSTM shows a worse performance. Though LSTM helps to address the gradient vanishing issue, for such 1 minute long spoken response, using the information passed to the last time step may still be not enough for accurate predictions
- The attention mechanism along with the BD-LSTM RNN model provides higher performance than the conventional model

System	Pearson r
Conventional model	0.585
CNN	0.581
BD-LSTM	0.531
BD-LSTM w/ attention	0.602

1. W. Hu, Y. Qian, F. K. Soong, 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.
2. J. Tao, S. Ghaffarzadegan, and L. Chen, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *ICASSP*, Shanghai, China, 2016.
3. D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic Text Scoring Using Neural Networks," in *Proc. of ACL*, June 2016.
4. Kaveh Taghipour and Hwee Tou Ng, "A neural approach to automated essay scoring.," in *EMNLP*, Sept. 2016
5. Yu et al. "Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech", in *ASRU*, Dec. 2015
6. Colin Raffel and Daniel P. W. Ellis, "Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems," *arXiv preprint arXiv:1512.08756*, Dec. 2015.