Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners

Angeliki Metallinou, Jian Cheng

Knowledge Technologies, Pearson
4040 Campbell Ave., Menlo Park, California 94025, USA
angeliki.metallinou@pearson.com, jian.cheng@pearson.com

10.21437/Interspeech.2014-358

Abstract

We investigated the use of context-dependent deep neural network hidden Markov models, or CD-DNN-HMMs, to improve speech recognition performance for a better assessment of children English language learners (ELLs). The ELL data used in the present study was obtained from a large language assessment project administered in schools in a U.S. state. Our DNN-based speech recognition system, built using rectified linear units (ReLU), greatly outperformed recognition accuracy of Gaussian mixture models (GMM)-HMMs, even when the latter models were trained with eight times more data. Large improvement was observed for cases of noisy and/or unclear responses, which are common in ELL children speech. We further explored the use of content and manner-of-speaking features, derived from the speech recognizer output, for estimating spoken English proficiency levels. Experimental results show that the DNN-based recognition approach achieved 31% relative WER reduction when compared to GMM-HMMs. This further improved the quality of the extracted features and final spoken English proficiency scores, and increased overall automatic assessment performance to the human performance level, for various open-ended spoken language tasks.

Index Terms: child speech recognition, deep neural networks, rectified linear units, English language learners, language proficiency assessment

1. Introduction

English language learners (ELLs) are students acquiring the English language for their education, and often coming from non-English-speaking homes or backgrounds (No Child Left Behind-NCLB,[1]). In the past decades, the percentage of ELLs in the US student population has greatly increased [2, 3]. Teachers face the challenge of how to better address the learning needs of these students, as some ELLs have difficulties in English usage and understanding, which may impede their academic achievement [4]. NCLB requires all states to identify ELLs and assess their English proficiency. The state of Arizona (with one of the highest ELL populations [3]) has partnered with Pearson to develop an automatic test for fast and consistent assessment of ELL students from kindergarten up to grade 12 (K-12). Automatic assessment could supplement teacher assessment, and could enable educators to devote more time teaching, and less time testing, in the classroom.

Past work on children's automatic assessment of oral reading fluency includes systems that score performance at the passage-level [5, 6, 7] or word-level [8]. A survey of spoken language technologies for education can be found in [9]. Here, we employ a series of open-ended spoken tasks to assess not only manner (pronunciation and fluency) but also understanding and usage of linguistic content. Automatic speech recognition (ASR) of children's open-ended speech is significantly more challenging than adult speech [10], partly because of the larger spectral and temporal variability in children speech which makes acoustic modeling more difficult [11]. In our ELL data, accented or mis-pronounced words, and student uncertainty about the task, often resulting in hesitations, mouth noises and unintelligible speech, represent additional challenges for our assessment system.

Recently, the use of deep learning for ASR has showed great promise [12]. Deep neural networks (DNNs) have been shown to work better than Gaussian mixture models (GMMs) for acoustic modeling [12, 13]. Various deep learning techniques, including restricted Boltzmann machine (RBM) pretraining [14] and rectified linear units [15, 16], are currently explored in the literature.

This work explores the use of DNNs for ELL children speech recognition. We investigate the use of DNNs to improve acoustic modeling of our challenging dataset that contains accented, child speech. We further explore how ASR improvements translate to better modeling of the linguistic content of students' responses through use of latent semantic analysis (LSA), and more accurate subsequent feature extraction for fluency and pronunciation modeling. We find that DNNs trained with rectifier activation functions greatly outperform GMMs, particularly for cases of mouth noises, hesitations and noisy responses (such cases are critical for our system). DNN acoustic models work better than GMMs, even when GMMs are trained with approximately eight times more data, which is encouraging, as large children speech datasets are often hard to collect. Overall, the use of DNNs leads to an average 31% relative WER reduction when GMM and DNN based systems are trained on the same data, and increases final assessment performance in terms of human-machine correlation from 0.795 to 0.826. This raises automatic performance at the human performance level, which validates the effectiveness of our assessment system, and outlines the value of deep learning for ASR-based children's education applications.

2. The AZELLA assessment system

The Arizona English Language Learner Assessment (AZELLA) is an ELL test, developed by Pearson and administered in Arizona K-12 students (described in detail in [17]). The speaking component of the test is delivered via speakerphone, and the student performance is automatically scored. The tests consist of a variety of spoken tasks which are developed by professional educators to elicit relatively open-ended displays of speaking ability. The tasks are summarized in Table 1. Responses are typically two to three sentences long.

Table 1: *The AZELLA spoken tasks.*

task	description
A	Answer question about an image
B	Give directions from a map
C	Ask question about an object
D	Answer open question on a topic
E	Give instructions for a task

This research focuses on scoring the pilot data obtained from elementary school students of ages 7-8 years old (a young and challenging age group). Those consist of 1500 spoken tests, each containing multiple responses, some belonging to the same task. 1200 tests are used for training, and 300 for testing, which results in 55 hours and 16 hours of training and testing data respectively. Student responses were transcribed and graded in terms of English proficiency level on a scale from 0-4, 4 being the highest level (for detailed grading rubrics see [17]). Professional human graders were recruited to provide a grade for each student response. Each response is double graded and the average is considered as the ground truth.

Our machine scoring system includes a combination of ASR, speech and text processing, and machine learning to capture both the linguistic content as well as the manner of speaking in order to assess the speaker's proficiency level. The system components, outlined in Figure 1, include ELL and native ASR systems, e.g., trained using ELL and native English data respectively. The speech recognition output is obtained from the ELL ASR, while the native ASR is used to compute spectral features through forced alignment (Sec. 4.2). Other components include content modeling using LSA and extraction of various spectral, duration and confidence score features, for fluency and pronunciation modeling. Task-specific neural network (NN) regression is used to compute the proficiency score.

3. CD-DNN-HMMs for speech recognition

3.1. CD-DNN-HMMs and pretraining

The CD-DNN-HMM framework for speech recognition, described in detail in e.g., [13], consists of replacing the Gaussian mixture models (GMMs) that were traditionally used for acoustic modeling, with deep neural networks (DNNs). A DNN is a multi-layer perceptron (MLP) with many hidden layers. Each layer computes the activations of conditionally independent hidden units given an input vector. If we denote as $\boldsymbol{u}_{l-1}$ the input of a hidden layer l (which is also the output of layer $l-1$), then the output of layer l can be computed as $\boldsymbol{u}_l = \sigma(\boldsymbol{W}_l \boldsymbol{u}_{l-1} + \boldsymbol{b}_l)$. $\boldsymbol{W}_l$ and $\boldsymbol{b}_l$ are the weights and biases of layer l, and σ is the network activation function, which is often chosen as the sigmoid function $\sigma = (1+\exp(-x))^{-1}$. The input of the first layer is the acoustic observation $\boldsymbol{u}_0 = \boldsymbol{o}(t)$, at time t, typically in the form of spectral features, e.g., MFCCs or Mel-scale log filterbank (FBANK) [18], computed over a speech window. Typically, a number of frames around the current frame are also included as context. The output layer uses the softmax function for each context-dependent phone state (senone) s, and the DNN estimates the posterior probability $P(s|\boldsymbol{o}(t))$ of each senone given the current observation:

$$P(s|\boldsymbol{o}(t)) = \frac{\exp(\boldsymbol{W}_l \boldsymbol{u}_{l-1} + \boldsymbol{b}_l)}{\sum_s \exp(\boldsymbol{W}_l \boldsymbol{u}_{l-1} + \boldsymbol{b}_l)} \quad (1)$$

Posteriors are then scaled by the prior probabilities of senones, e.g., $P(s|\boldsymbol{o}(t))/P(s)$, to obtain the scaled likelihoods, as required by the decoding process [19]. In our experiments, this scaling is important, especially for silence posteriors, since our data contains long silences (responses are preprocessed to automatically detect and remove leading and trailing silences, however long pauses between words are still common).

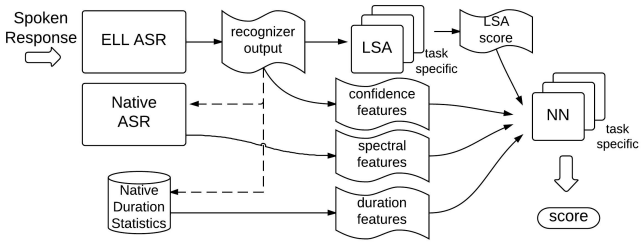

Figure 1: Outline of the AZELLA speech recognition and language assessment system.

Stochastic gradient descent and error back-propagation are typically used for learning the DNN, while the senone labels of the training speech frames are obtained through forced alignment, using an already trained model, e.g., a CD-GMM-HMM. However, deep networks might get stuck in poor local optima during training when network weights are initialized randomly. An unsupervised method for model pretraining (initialization) has been proposed in [14], which uses RBMs. RBM pretraining is widely used in the literature [13, 20], while some researchers have argued that, although it often helps, it is not necessary for training good models [21].

3.2. Rectified linear units

Researchers have suggested the use of the activation function $\sigma = max(0, x)$ for the hidden layers of a DNN, in which case the hidden units are commonly called rectified linear units (ReLUs) [15, 22]. CD-DNN-HMMs with ReLUs have been successfully applied for speech recognition, and their advantages have been attributed, among others, to better regularizing the internal network representation by producing many zero activations during training [16]. In our experiments, using ReLUs allows us to easily train deeper networks faster than when using sigmoids, and without pretraining (which agrees with [16]).

4. English proficiency assessment

4.1. Content scoring based on LSA

Content scoring focuses on whether the student is able to respond to the task with appropriate linguistic material, and for this purpose, we use a method similar to Latent Semantic Analysis (LSA) [23]. LSA builds a co-occurrence matrix of words and their usage in documents (here responses), and then reduces the matrix using Singular Value Decomposition (SVD). The similarity between two words (rows of the original matrix) can be measured in this reduced space.

Here, we collect training responses with good human scores, compute word and word sequence occurrences, and use SVD to create a space of 'good' responses. A test response is then scored by scaling the weighted sum of the occurrence of a large set of words and word sequences (expected in 'good' responses) that may be recognized in the test response. Weights are assigned to the expected words and word sequences according to their semantic relation to 'good' responses using a method similar to LSA. Intuitively, test responses that are more similar to the 'good' training responses will receive a higher

score. We train LSA models specific to each task (by design, the five different tasks are roughly equally represented in our train and test dataset).

4.2. Spectral, duration and confidence features

Proficiency scoring includes not only linguistic content but also manner scoring, e.g., fluency and pronunciation. To incorporate these attributes to our final score, we further extract confidence, spectral and duration features.

Spectral: To consider manner (pronunciation and fluency), we included spectral likelihood features, that are computed by comparing the recognition log likelihoods between native and ELL ASR. Specifically, we use the native ASR to perform forced alignment of the recognition output obtained from the ELL ASR, and collect the log likelihoods lp_i^{fa} for each phoneme p_i. Then, given the computed phoneme boundaries, we perform allphone recognition, i.e., we let the native ASR pick the most likely phoneme for a given segment (of duration d_i), with corresponding likelihood lp_i^{ap}. If S is the set of N phonemes, we compute the average spectral score difference as:

$$spectra = \frac{1}{N} \sum_{i \in S} \frac{lp_i^{fa} - lp_i^{ap}}{d_i}$$

In practice, we compute feature variations using different phoneme sets S, e.g., predefined phonemes that tend to be challenging for students. Spectral features are inspired by common methods for pronunciation assessment [24, 25], and were validated in [26]. As an extra feature, we include the percentage of phonemes from the allphone recognition matching the phonemes from the force alignment.

Duration: Duration features compare duration between the student's speech (non-native) and native speech statistics (some materials are adopted from our past work [27]). Specifically, we use a large corpus of native English children data [28] to precompute phoneme duration statistics. Then, for a sequence of N recognized phonemes p_i (of duration d_i) in the student's non-native response, we compute the average log probability $log\_pr = 1/N \sum(log(P(d_i)))$, where $P(d_i)$ is the probability that the observed duration of phone p_i could have been produced by a native speaker. We also compute a similar log probability feature for interword silence durations. Note that such features are empirically found not to be very sensitive to speaking rate: our preliminary experiments with speaking rate normalization did not bring substantial performance change.

Confidence: The ASR module assigns confidence scores to the recognized words and phonemes [29]. For each response, we compute the percentage of words and percentage of phonemes whose confidence is lower than a (experimentally defined) threshold, as features to predict student performance.

We use the above features, along with the LSA scores, as input to NN regression, in order to compute the final proficiency score. We train NN regression models specific to each task.

5. Experiments and results

5.1. Speech recognition results

Our CD-GMM-HMM system uses an augmented ASR based on HTK [30]. For our CD-DNN-HMM system, we augmented the TNet toolkit [31, 32] and modified HTK to accept the DNN acoustic models. To investigate the effect of variable amount of data, we use two training sets: tr_1 which contains the 1200 elementary student tests from the AZELLA dataset (55 hours), and tr_2 which additionally contains data from other AZELLA test stages (primary, middle and high school) and from other children datasets [6] available to Pearson (430 hours). Our testing set contains 300 elementary student tests (16 hours) from AZELLA, that are not in tr_1 or tr_2. Here, we focus on the ELL acoustic models, since those are used for recognition (Figure 1).

Table 2: *Baseline CD-GMM-HMM recognition performance, trained on different amounts of data.*

training set	WER
tr_1 (55h)	29.7
tr_2 (430h)	27.8

For the GMM acoustic features, we use a standard configuration of 12 MFCCs plus energy, with first and second derivatives, extracted with window 25 msec and frame rate 10 msec. The DNN acoustic features are 40 Mel-scale log filterbank (FBANK) plus energy, with first and second derivatives, extracted with window 25 msec and frame rate 10 msec. This choice of features is based on previous work [20, 12, 33], that suggests that FBANK features may work better than traditional MFCCs for DNN-HMM. For the DNNs, we also concatenate a context of eight feature frames left and right of the current frame, i.e., 17 frames total. DNN models consist of hidden layers with 2K units. The above DNN configuration is empirically found to work well. Finally, we use task-specific training data to build language models constrained to each of the five spoken tasks of Table 1, to be used during recognition.

Table 2 shows the CD-GMM-HMM baseline performance when trained on tr_1 and tr_2. The number of mixtures were empirically optimized to 32. As expected, the GMM benefits from the use of a large training dataset. Table 3 shows the CD-DNN-HMM system results, when varying the number of hidden layers, the initialization method (RBM pretraining or random) and the activation function (sigmoid or ReLU) for the DNN model. The DNNs were trained on the smaller tr_1 set and contain about 1K output senones. According to our results, deeper networks work better, and 5-layers seems a good configuration. DNNs with ReLUs outperform DNNs with sigmoid units, even when the latter are initialized with RBM pretraining. To our experience, using ReLUs enables training deeper networks fast and without much parameter tuning. Networks with sigmoid units are harder to converge to good weight values, and they require more careful experimentation with initial parameters (e.g., learning rate, magnitude of initial random weights). RBM pretraining may slightly improve performance and reduces the parameter tuning effort, especially when training deeper networks, but adds a significant computational cost. From Table 3, we notice that DNNs significantly outperform GMMs trained on the same amount of data (first line of Table 2), and most times work better than GMMs trained on approximately eight times more data (second line of Table 2). The best configuration is a 5 layer ReLU DNN.

Table 3: *CD-DNN-HMM recognition performance (WER) for various DNN configurations and training strategies. DNNs are trained on tr_1 (55h).*

number of hidden layers	sigmoid random init.	sigmoid RBM pretrain	ReLU random init.
2	27.1	27.8	26.2
3	26.4	26.8	25.9
4	27.0	27.1	26.1
5	27.8	27.1	**25.5**
6	27.3	26.1	27.2

Choosing the optimal DNN configuration from Table 3, we use the larger train set tr_2 to train another 5 layer DNN, containing about 3K output senones. To better understand the behavior of our models, we focus on cases where the student seems to be having difficulty to respond. Such cases are commonly accompanied with more mouth and breathing noises, hesitations, partial words, low volume speech, unclear and sometimes unintelligible speech, and are challenging for our ASR system. The above cases are not transcribed in detail, however we can approximately select them by choosing student responses containing less than three words. Typical responses should contain two to three sentences, and to our experience, extremely short responses often represent problematic cases. We split our test set into two subsets, one containing responses shorter than three words (named $Shorter$ and containing approximately 30% of all responses), and the other containing the remaining responses ($Longer$).

The recognition results on each subset can be seen in Table 4. Looking at the $Shorter$ dataset results, the GMM didn't handle well these challenging cases. A common mistake is misclassifying mouth/breathing noises, hesitations and other vocalizations as speech, possibly because of the large variability of the acoustic models for child speech. DNNs seem to handle much better this variability and perform reasonably well, even when trained on the smaller set tr_1. This agrees with findings in [34], that suggest DNNs are robust to variability in the speech signal, as long as they are trained on a sufficiently representative training set. Overall, comparing GMM and DNN performance over the total test set (All) when models are trained using tr_2, we have achieved a 31% relative WER reduction of the DNN system (27.8% vs 19.3%).

Table 4: *Comparison of WER of the CD-GMM-HMM and CD-DNN-HMM systems for different subsets of the test data. All denotes the full test set, $Shorter$ contains responses with fewer than 3 words and $Longer$ contains the remaining responses. Training set tr_1 contains 55h data and tr_2 contains 430h data*

model	train set	test set		
		All	*Shorter*	*Longer*
GMM (32 mix)	tr_2	27.8	78.6	26.4
DNN (5 layer ReLU)	tr_1	25.5	51.7	24.9
DNN (5 layer ReLU)	tr_2	**19.3**	**41.3**	**18.8**

5.2. Proficiency assessment results

We explore the advantage that our improved ASR may give in better estimating proficiency scores. We compare two systems, both trained on tr_2: the CD-DNN-HMM system (5 layer ReLU DNN) and the CD-GMM-HMM baseline (32mix GMM), in terms of the final assessment performance they achieve. For these results, we additionally trained a native CD-DNN-HMM system (5 layer ReLU, trained on native children data), which is used for forced alignment during spectral feature extraction as explained in Section 4.2. Our observation is that system performance mostly depends on the ELL CD-DNN-HMM system which is used for recognition (using GMM or DNN for the native system does not significantly alter performance).

We use the ELL CD-DNN-HMM system to decode all 1500 train and test responses. Then the train response recognition output is used to build task-specific LSA models, as explained in Section 4.1. The LSA models are then used to score the test responses in terms of their linguistic content (LSA score). We similarly train baseline LSA models, using the CD-GMM-HMM system. In Table 5 (columns 2-3), we present the correlation between LSA scores and the average human grades (H), for both systems. We notice that the use of a CD-DNN-HMM system improves correlations for all tasks, while the largest performance gains are for tasks B, C and E. Overall correlation improves from 0.758 to 0.801. As expected, better ASR leads to better scoring of the linguistic content of the test responses.

Table 5: *Comparison of scoring linguistic content and overall proficiency between the GMM-based and the DNN-based system, both trained on tr_2. Performance is measured by the correlation between machine estimated score (either LSA or final score) and the average human grade (denoted as H). Human-human correlation is also presented for comparison.*

task	corr. LSA-H		corr. final score-H		H-H
	GMM	DNN	GMM	DNN	
A	0.811	0.822	0.865	0.871	0.874
B	0.827	0.867	0.844	0.879	0.819
C	0.624	0.705	0.702	0.762	0.831
D	0.738	0.758	0.746	0.772	0.753
E	0.793	0.850	0.818	0.845	0.834
total	0.758	0.801	0.795	0.826	0.822

Next, for each system, we use the training data to build a NN regression model for scoring. The input features are the computed LSA score and the spectral, duration and confidence features described in Section 4.2. The correlation between the final scores and the average human grades is presented in Table 5 for both systems (columns 4-5). The human-human correlation is also presented for comparison reasons (column 6). Again, the use of DNNs, leads to higher correlations for all spoken tasks. For certain tasks, we observe large improvement, for example task C correlations increase from 0.702 to 0.762. Overall correlation increased from 0.795 to 0.826, which is a 0.031 absolute improvement. Our final human-machine corelation is at the same level as human grader correlation (0.822 overall). This validates the effectiveness of our proficiency scoring system, and outlines the usefulness of deep learning techniques for building ASR-based children's applications.

6. Conclusions and future work

In this work, we have explored the use of deep learning techniques for increasing ASR accuracy for better proficiency assessment in children's educational applications. We worked on a dataset of ELL children's speech, collected from students interacting with Pearson's automatic assessment system, which contains various technical challenges including accented speech, background noise, mouth and nonverbal noises, hesitations and others. Based on our experiments, using a CD-DNN-HMM system for ASR, with a 5 layer ReLU DNN, greatly ourpeforms the traditional GMM-based system, and leads to a 31% relative WER reduction. Large improvement was specifically observed for the above challenging cases of noisy and unclear responses. This further improves the quality of our speaking proficiency features, which describe linguistic content, fluency and pronunciation of the student's responses, and leads to an improvement of overall language proficiency assessment in terms of human-machine corelations. We achieve overall automatic assessment performance that is at the human performance level, which validates the effectiveness of our system.

In the future, we plan to explore the advantages of deep learning techniques for other challenging datasets of non-native speech in various languages, and incorporate such techniques in Pearson's series of educational language assessment products.

7. References

[1] "The No Child Left Behind Act (NCLB)," Public Law No. 107-110, 115 Stat. 1425, 2002.

[2] K. Flynn and J. Hill, "English language learners: A growing population (policy brief)," Aurora, CO: Mid-continent Research for Education and Learning, Tech. Rep., 2005.

[3] R. Payan and M. Nettles, "Current state of English language learners in the U.S. K-12 student population," Princeton, NJ: ETS., Tech. Rep., 2006.

[4] P. Gandara, J. Maxwell-Jolly, and A. Driscoll, "Listening to teachers of English language learners: A survey of California teachers' challenges, experiences, and professional development needs," *Policy Analysis for California Education, PACE (NJ1)*, 2005.

[5] K. Zechner, J. Sabatini, and L. Chen, "Automatic scoring of childrens read-aloud text passages and word lists." in *Proc. of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, 2009.

[6] J. Cheng and J. Shen, "Towards accurate recognition for children's oral reading fluency," in *IEEE-SLT 2010*.

[7] R. Downey, D. Rubin, J. Cheng, and J. Bernstein, "Performance of automated scoring for children's oral reading," in *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, 2011.

[8] J. Tepperman, M. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan, and S. Narayanan, "A Bayesian network classifier for word-level reading assessment," in *Proc. of ICSLP*, 2007.

[9] M. Eskanazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, 2009.

[10] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Proc. of Speech and Language Technologies for Education (SLaTE)*, 2007.

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

[12] G. Hinton, L. Deng, Y. Dong, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, 2012.

[13] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio Processing, Special Issue on Deep Learning for Speech and Lang. Processing*, 2012.

[14] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, 2002.

[15] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. of ICML*, 2010.

[16] M. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. Hinton, "On rectified linear units for speech processing," in *Proc. of ICASSP*, 2013.

[17] J. Cheng, Y. Zhao-D'Antilio, X. Chen, and J. Bernstein, "Automatic spoken assessment of young English language learners," in *Proc. of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[18] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. of ICASSP*, 2012.

[19] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[20] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, 2012.

[21] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures (section 3.2)," arXiv, Tech. Rep., Sept. 2012.

[22] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *J. of Machine Learning Research*, vol. 15, 2011.

[23] T. Landauer, P. Foltz, and D. Laham, "Introduction to latent semantic analysis," *Discourse Processes*, vol. 25, 1998.

[24] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Proc. of ICASSP*, 1997.

[25] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, 1999.

[26] J. Bernstein, A. V. Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, 2010.

[27] J. Cheng, "Automatic assessment of prosody in high-stakes English tests," in *Proc. of Interspeech*, 2011.

[28] J. Bernstein and J. Cheng, *Logic and validation of a fully automatic spoken English test*, V. M. Holland and F. P. Fisher, Eds. The Path of Speech Technologies in Computer Assisted Language Learning, New York: Routledge, 2007.

[29] J. Cheng and J. Shen, "Off-topic detection in automated speech assessment applications," in *Proc. of Interspeech*, 2011.

[30] S. J. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book, version 3.4*. Cambridge University Engineering Department, 2006.

[31] "Neural Network Trainer TNet." [Online]. Available: http://speech.fit.vutbr.cz/software/neural-network-trainer-tnet

[32] K. Vesely, L. Burget, and F. Grezl, "Parallel training of neural networks for speech recognition," in *Proc. of Interspeech*, 2010.

[33] L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. L. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at Microsoft," in *Proc. of ICASSP*, 2013.

[34] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition tasks," in *Proc. of ICLR*, 2013.