INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

DNN Online with iVectors Acoustic Modeling and Doc2Vec Distributed Representations for Improving Automated Speech Scoring

Jidong Tao, Lei Chen, Chong Min Lee

Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
{jtao, lchen, clee001}@ets.org

Abstract

When applying automated speech-scoring technology to the rating of globally administered real assessments, there are several practical challenges: (a) ASR accuracy on non-native spontaneous speech is generally low; (b) due to the data mismatch between an ASR systems training stage and its final usage, the recognition accuracy obtained in practice is even lower; (c) content-relevance was not widely used in the scoring models in operation due to various technical and logistical issues. For this paper, an ASR in a deep neural network (DNN) architecture of multi-splice with iVectors was trained and resulted in a performance at 19.1% word error rate (WER). Secondly, we applied language model (LM) adaptation for the prompts that were not covered in ASR training by using the spoken responses acquired from previous operational tests, and we were able to reduce the relative WER by more than 8%. The boosted ASR performance improves the scoring performance without any extra human annotation cost. Finally, the developed ASR system allowed us to apply content features in practice. Besides the conventional frequency-based approach, content vector analysis (CVA), we also explored distributed representations with Doc2Vec and found an improvement on content measurement.

Index Terms: automated speech scoring, non-native spontaneous speech, automatic speech recognition, unsupervised language model adaptation, content vector analysis, Doc2Vec

1. Introduction

The application of modern speech processing technology in the field of automated speech assessment [1] has been intensively studied for the past two decades. Automatic speech recognition (ASR) is generally used to provide the information needed for extracting a variety of features covering many speaking abilities, e.g., delivery, language use, and topic development in the TOEFL® internet-based test (iBT) [2]. Based on these computed features, statistical models were trained to automatically predict language learners' speaking proficiency levels [3, 4].

As suggested in [5], the ASR module inside a speech scoring system plays important roles in the system's overall scoring performance, and the ASR's recognition accuracy matters more for the measurements focusing on high-level linguistic performance, e.g., content relevance. However, compared to the ASR performance on native speech data, the ASR task on non-native language learner's speech data is more challenging, and the corresponding recognition accuracies are relatively low. For the automated scoring of globally administered speech assessments, a data mismatch pattern always exists [6]. This pattern refers to the fact that the speech data collected in real operational assessments could be different from the data used in the ASR training stage in various respects, e.g., speaker profiles, recording environment, speaking content, etc. Clearly, such a mismatch further diminishes the ASR's accuracy on the real assessment data collected. Another factor impacting the performance of automated speech scoring systems is that content measurement has not been commonly used. The lack of content measurement was caused by technology and logistic reasons. From the technology side, it is an interesting research question to obtain a proper representation (of content) based on a spoken response's ASR output, which tends to be noisy given unavoidable ASR errors. From the logistical side, for any assessment with many different questions, training question-specific topic models is demanding in labor and cost.

In this paper, we will report on our efforts to overcome these challenges that impact the automated scoring of a real large-scale speech assessment, including (a) a highly accurate DNN ASR targeting non-native spontaneous English from learners around the world; (b) an unsupervised learning framework applying the language model (LM) interpolation method to improve both recognition accuracy and scoring performance by utilizing historical assessment data, i.e., spoken responses and human rated scores from a different assessment; (c) the investigation of using the newly emerging Doc2Vec [7] technology to generate vector representations on spoken responses' ASR hypotheses, and (d) the evaluation of the inclusion of content-related features in the speech scoring task.

The paper is organized as follows: Section 2 briefly reviews related previous research; Section 3 describes the English tests, the datasets for system training and evaluation, and the research questions we try to answer in this paper; Section 4 describes the ASR-related research in this study; Section 5 describes using vector space modeling to generate the content-related features; Section 6 reports our experimental results; and finally Section 7 provides conclusions and directions for future work.

2. Previous work

An automated speech assessment system typically takes the output of an ASR to generate the features for scoring. A working example of a rich set of speech features can be found in [3]. With deep learning technology's increasing prominence in ASR, it has also recently been applied to the task of speech assessment [8, 9, 10]. The more recent work [11] reported a word error rate (WER) of 22.76% as the best ASR performance using a DNN acoustic model trained on an approximately 800-hour large-vocabulary non-native spontaneous English corpus.

http://dx.doi.org/10.21437/Interspeech.2016-1457

Adaptation is an effective method to improve ASR performance by reducing model mismatch when applying a trained model to out-of-domain data [12]. Out-of-domain data can differ from training data in several ways, including different types of speakers, audio qualities, speaking content, etc. In particular, speaker adaptation refers to the adaptations applied on the dimension of acoustic information, including updating the acoustic models on either the feature level or the model level. For example, *iVector*, a dominant method in the area of voice biometrics to represent both speaker and channel variability in low-dimensional vector space, has been commonly applied as a speaker adaptation approach in the field of ASR to improve recognition accuracy [13, 14, 15]. One of the standard approaches in LM adaptation is to use linear interpolation to mix a LM trained from an out-of-domain dataset with a generic LM in order to improve the final LM's coverage. When it is not practical to obtain manual transcriptions of the out-of-domain speech data, unsupervised LM adaptation, which uses an existing ASR to recognize the out-of-domain speech data to generate approximate transcriptions, can be used. [16] showed how this technique can be used to improve ASR in automated speech scoring.

With more accurate ASR systems on test-takers' non-native speech data, the research was initiated to investigate the performance of different types of content features using NLP technologies. [17] described the application of latent semantic analysis (LSA) to automated essay scoring. [18] explored the use of a range of content similarity features in the context of an assessment of spoken English for academic purposes, including LSA, pointwise mutual information (PMI), and cosine similarity based on content vector analysis (CVA). In these previous efforts to include content measurement into the task of automated scoring [19, 18], the CVA method based on vector space modeling has gotten a lot of attention.

The above mentioned approaches use single numeric value to represent each word token. For example, TF-IDF representation is used in CVA. However, since 2011, with the introduction of **Word2Vec** [20], distributed representation of words has gotten substantially increasing attention due to this several technical advantages and a series of successful applications in diverse NLP tasks. **Doc2Vec** [7] further extends the distributed representation from the word level to the document/paragraph level. The learned Doc2Vec vectors can be directly used for document classification, e.g., sentiment detection [7]. In particular, the Word2Vec method produces the word vectors from an input text corpus. It first constructs a vocabulary from the training text data and then learns vector representations of words. The Word2Vec representations can be learned by two different methods: continuous bag of words (CBOW) and skip-gram. In the CBOW method, the goal is to predict a word given the surrounding words, whereas the goal of the skip-gram method is to predict neighbor words given a single word. Quite similar to Word2Vec, Doc2Vec also uses an unsupervised learning method to learn the distributed word representations from a corpus. However, Doc2Vec introduces an additional vector to represent the entire paragraph/document's semantic information during the training process. The obtained vectors related to paragraph/document can be used to represent the paragraph/document.

3. Task and data

In this paper, we investigate the automated speech scoring technology used for rating the TOEFL Practice Online® (TPO), which is the official practice test for test takers to prepare the TOEFL iBT® test. The TOEFL iBT is a well known English test in the TOEFL® family, measuring test takers' readiness for attending universities using English as their primary instructional language. The TPO test has the following features: (a) it contains "retired" TOEFL iBT questions that won't be used in future operational tests, (b) it uses a different test delivery interface to allow TPO users to take tests in their homes using their computers and audio recording devices, and (c) it is fully automatically scored.

The ASR engine used in the TPO automated speech scoring system was trained from a large-sized TOEFL iBT transcribed corpus (more details can be found in Section 4). Prompts (a.k.a. question types) in this ASR training set are limited and cannot cover new prompts that have recently been retired from the TOEFL iBT prompt bank. Therefore, for more accurate scoring, it is necessary to do a LM adaptation to improve the ASR system's performance on these new prompts. Also, as mentioned in Section 2, applying content measurement requires a collection of scored responses for training content reference vectors. These continuously increasing new prompts means frequent human rating needs. However, adding these manual transcription and scoring tasks would substantially increase the cost of developing the assessment. Therefore, in this paper, leveraging the fact that all of the TPO prompts are based on the TOEFL iBT prompts, which have previously been used with adequate speech responses and human rated scores, we are interested in answering the following research questions (RQ), including:

RQ1: Can we use unsupervised LM adaptation and the spoken responses from the TOEFL iBT test's previous operations to improve ASR performance on the TPO prompts that are not covered in ASR training?

RQ2: Can the new distributed representation, i.e., Doc2Vec, play a useful role in content scoring?

RQ3: Can we use the TOEFL iBT historical data, i.e., human rated scores and ASR hypotheses on the spoken responses, to conduct content relevance measurement in the TPO test without any extra human labeling cost?

To answer these questions, we prepared the datasets for our experiments as follows: we obtained TPO spoken responses from 24 prompts that were not covered in ASR training; these responses were divided for scoring model training (sm-train), which consisted of 2089 responses from 355 speakers, and for scoring model evaluation (sm-eval), which consisted of 851 responses from 149 speakers. All responses used for scoring were double scored by experienced human raters following the 4-point scale scoring rubric designed for scoring the TOEFL iBT [1]. The scoring reliability is measured by the inter-rater agreement calculated in terms of both the Pearson correlation coefficient (r) and quadratic weighted kappa (κ). The item (a.k.a. response) level and the speaker level (i.e. the sum of 6 item level scores) inter-rater agreements are r_{item} = 0.59, κ_{item} = 0.58, and r_{spkr} = 0.87, κ_{spkr} = 0.86 respectively. Then, for these 24 prompts, we acquired spoken responses and human-rated scores from the archive of historical TOEFL iBT data. For each prompt, 1000 responses with different score levels (1 to 4) were sampled. This dataset will be called *the historical data* in the remainder of the paper.

[1] https://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

4. ASR systems

In [11], the best ASR performance in terms of word error rate (WER) was 22.76% when deep neural networks (DNNs) were used for acoustic modeling on the training partition that contains a total of 819 hours of non-native spontaneous speech covering more than 100 L1s across 8700 speakers from about 150 countries around the world. For this study, the same ASR training corpus as in [11] was used to build a highly improved acoustic model (AM) using the DNN multi-splice online training with iVectors.

In particular, the new AM was built by following [21]. In this system, a 6-layer DNN with *p*-norm (*p*=2) nonlinearity is trained using layer-wise supervised back-propagation training [22]. Frames of 13-dimensional Mel-frequency cepstral coefficients (MFCCs) along with their Δ and $\Delta\Delta$ coefficients are extracted as acoustic features using a 25ms frame-size with a 10ms shift for 16kHz 16-bit mono wav files. An iVector, a vector of 100 dimensions per frame which represent speaker properties, is appended to the MFCCs together as input to the DNN training. iVectors for speakers are estimated in an online mode, where the frames prior to the current frame, including the previous utterances of the same speaker, are used. The DNN does milti-splicing temporal windows of frames over time at each layer, and a sub-sampling technique is used to reduce computational cost. [21] indicated that it works better to start out splicing the context frames over close-together frames (e.g., 2, 1, 0, 1, 2), and splice over further-apart frames (e.g., -7, 7) in deeper layers. A normalization component right after each hidden layer is applied to keep the training stable and prevent the neurons from becoming "over-saturated". To achieve the best performance, sequence-discriminative training based on a state-level variant of the minimum phone error (MPE) criterion, called sMBR [23], is applied on top of the DNN. A trigram statistical LM with about $525K$ tri-grams, $605K$ bi-grams over a lexicon of $23K$ words in this ASR system was trained using modified Knesser-Ney discounting [24] by SRILM [25] on the manual transcriptions of the same AM training partition, which consists of $5.8M$ word tokens. This LM serves as the generic LM.

Using this improved ASR system built from the ASR training corpus, we decoded the $24,000$ spoken responses in our historical dataset to train a domain-specific LM. Then, the generic and domain-specific LMs are linearly interpolated to build the adapted LM. The interpolation weight λ was decided by the perplexities between the generic and domain-specific LMs. Note that applying speaker adaptation on ASR systems using DNN AMs is not a trivial task. Several well-performing speaker adaptation methods in a Gaussian Mixture Model (GMM) AM based ASR, such as maximum a posterior (MAP) [26], and maximum likelihood linear regression (MLLR) [27], can not be instantly used in DNNs. Also, training a data domain specific DNN from scratch is empirically recommended to achieve a reliable ASR performance, supposing the training data is sufficient.

5. Vector space for measuring content

Regarding measuring content relevance, as suggested in [19, 18], the CVA method is widely used. In a CVA model, a spoken response's ASR output is firstly converted to a vector. Each cell in the vector is a word's term frequency (TF) normalized by the inverse document frequency (idf), called to be tf-idf thereafter. The content relevance between two responses can be measured as the distance (e.g., cosine similarity score) between the two vectors. Typically, for each score level, a reference vector was trained using a set of responses that had that score. Then, for an input response, the distances between its corresponding vector and these reference vectors are used as content features. For each response's vectorization plan, five features can be extracted. cos_i refers to the cosine similarity between the input response's vector and a score-level (1 to 4) reference vector. $argmax_{cos}$ refers to the score level judged by the maximum cosine similarities.

Doc2Vec is a new approach in NLP to obtain vectors. Vectors for each response are produced in a configuration similar to that in [7], in which the each vector contains 100 elements. Distributed Memory (DM) and Distributed Bag of Words (DBOW) are two primary training methods for obtaining Doc2Vec representations. DM can be further grouped into DMC and DMM. The difference between DMC and DMM is that the former concatenates context vectors, whereas the latter averages them. DMC consumes more memory during training, and results in a larger model. The numbers of context words surrounding the the predicted word are 5 and 10 for DMC and DMM respectively. DBOW forces the model to predict groups of words randomly sampled from the given vector. In practice, DBOW and DM models can be combined together to provide other types of vectors, such as DBOW+DMC, and DBOW+DMM. In our case, we want to build vectors of the adaptation data to represent each of the score points. As for training the reference vectors by using the Doc2Vec vectorization approach, we train a set of individual Doc2Vec vectors from a set of responses for a particular score level. Then, the mean vector from all of these vectors are formed to be the reference Doc2Vec vector for this score level. By using the three different Doc2Vec training methods described above, we explored the five types of vectorization approaches as follows:

1. DMC: Distributed Memory model in Concatenating context vectors
2. DMM: Distributed Memory model in taking the Mean of context vectors
3. DBOW: Distributed Bag of Words model
4. DBOW+DMC: DBOW and DMC in concatenation
5. DBOW+DMM: DBOW and DMM in concatenation

Using the historical dataset ($24,000$ iBT responses), we trained reference vectors for the four score levels (1 to 4) by using the vectorization approaches described in this section, including using tf-idf values and the five types of Doc2Vec methods that are implemented in the Gensim Python package [28].

6. Experiments

Our new ASR system, built with the DNN multi-splice online AM with iVectors and a generic trigram LM, achieves a 19.1% WER on the asr-eval dataset, which is a 16% relative WER reduction compared to the DNN ASR reported in [11] using the same training and evaluation datasets. In fact, the performance of this system is close to human experts' WER of about 15% for non-native spontaneous speech, as reported in [29]. To our knowledge, this is the lowest WER reported on TOEFL iBT non-native spontaneous speech. This new ASR system provides more accurate ASR hypotheses for the unsupervised LM adaptation. Table 1 compares the ASR performance using the generic LM with using the adapted LM. Because the prompts in the scoring corpus have no overlap with those in the ASR training corpus, the ASR using the generic LM has WERs of 40.09%

and 38.84% on the sm-train and sm-eval partitions. Using unsupervised LM adaptation helps to reduce the WERs to 36.68% and 35.42% respectively, which are about 8.51% and 8.81% relative WER reductions. More importantly, this considerable WER reduction was achieved without any transcription cost.

ASR: DNN AM	$WER_{sm-train}$	$WER_{sm-eval}$
+ generic LM	40.09	38.84
+ adapted LM	36.68	35.42

Table 1: Unsupervised LM adaptation helps to reduce WER for the TPO speech responses whose prompts were not included in the ASR training data

Next, we addressed RQ2 to compare the two different document-to-vector approaches to measuring content relevance. On the sm-train dataset, we computed the Pearson correlations rs between the human-rated scores and the two types of vector space based content features, i.e., cos_4 and $argmax_{cos}$. A high r suggests that the corresponding feature is more predictive. Table 2 reports on the obtained r values using the tf-idf (in CVA) and five Doc2Vec approaches for forming vectors. Clearly, several Doc2Vec training approaches generate more indicative content measurement features. The $argmax_{cos}$ feature is chosen to score because it has a consistently higher correlation with human scores than cos_4 across the all methods.

Representation	cos_4	$argmax_{cos}$
CVA	0.286	0.390
DMC	0.332	0.339
DMM	0.283	0.382
DBOW	0.299	0.432
DBOW+DMC	0.314	0.418
DBOW+DMM	0.288	0.403

Table 2: Correlations of individual content features with human scores across 1 CVA system and 5 Doc2Vec systems.

Finally, we address RQ3 to carry out our ultimate evaluation task in this paper: evaluating the effects of the more accurate ASR and better content measurement on the speech scoring task. SpeechRater^{SM}, an automated scoring engine for assessing non-native English proficiency [3], is used to extract scoring features and predict a numerical score for spoken responses. The features, summarized in Table 1 in [11], are related to several aspects of the speaking construct[2], which include *fluency*, *rhythm*, *intonation & stress*, *pronunciation*, *grammar*, and *vocabulary use*. Automatic scoring feature selection based on LASSO regression [30] is used to obtain a much smaller input feature set for building a linear regression model for score prediction. Note that linear regression (LR) is used (instead of other more powerful machine learning algorithms) to obtain a more interpretable model.

Table 3 reports on our machine scoring experiment using the two trained ASR systems with different scoring features. When using the ASR system after the unsupervised LM adaptation, we can find that the scoring performance is improved, compared to the ASR system using the generic LM. Note that the LASSO regression selected a different number of features (#F), 32 vs. 28. After adding the all 5 Doc2Vec $argmax_{cos}$ features to the scoring model described in the second row, the scoring performance was further improved. Later, when adding one more $argmax_{cos}$ feature using the tf-idf CVA model, the overall scoring performance reached the highest level. In a summary, compared the result reported in the first row, which uses the ASR with a generic LM and lacks content measurement features, the final scoring model containing all of the investigated methods in this paper, has a considerable performance gain. In particular, on the item level, κ increases from 0.49 to 0.53, and on the speaker level, κ has increased from 0.77 to 0.80. The performance becomes closer to human-to-human agreement results. For example, the final model's r_{item} becomes very close to human-to-human (H-H) performance, 0.58 vs. 0.59.

[2]In psychometric terms, a *construct* is a set of knowledge, skills, and abilities that are required in a given domain.

System	#F	r_{item}	κ_{item}	r_{spk}	κ_{spk}
H-H		0.59	0.58	0.87	0.86
Generic LM	32	0.53	0.49	0.79	0.77
Adapted LM	28	0.54	0.50	0.80	0.77
+ D2V	33	0.56	0.51	0.80	0.78
+ CVA	34	0.58	0.53	0.82	0.80

Table 3: Pearson correlation (r) and quadratic weighted kappa (κ) between SpeechRater^{SM} and human raters' scores in item and speaker level across the two ASR systems with different scoring features.

7. Conclusions and future work

In this paper, regarding several technical challenges during the utilization of ASR-based speech scoring to support a large-scale assessment, we proposed our solutions. Our achievements can be summarized as follows:

- On top of a state-of-the-art DNN ASR developed on non-native spontaneous speech [11], we applied DNN online training with iVectors-based speaker adaptation to further reduce the ASR system's WER by 16.0% relatively.
- For the TPO spoken responses answering the prompts that were not included in the ASR training, we utilized a unsupervised LM adaptation to improve the LM's coverage and reduce the ASR WER by about 8.0%. Though having historical data (from a relevant but different assessment) is unique to our case, the same idea could be widely applied.
- To our knowledge, this is the first work utilizing distributed representation (Doc2Vec) rather than tf-idf based vectors in the vector space content measurement. Doc2Vec content features show higher predictive power than conventional tf-idf content features.

The research proposed in this paper helps us improve the overall scoring performance considerably and nearly reaches human-to-human agreement level, a gold standard in the automated assessment area. The experiments conducted in this paper provide solid answers to the three research questions we raised at the outset. Moreover, the proposed methods do not need any extra human transcription and scoring effort and are thus quite attractive from a economical point view.

Regarding future work, one direction will be continuously improving ASR performance on non-native speech. Given the many recent advance methods in ASR, e.g., end-to-end neural network, it is highly possible that ASR accuracy will completely reach human transcribers' level. Another direction will be refining the current implementation of Doc2Vec, and exploring its broader applications in automated assessment.

8. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] D. Higgins, L. Chen, K. Zechner, K. Evanini, and S.-Y. Yoon, "The impact of asr accuracy on the performance of an automated scoring engine for spoken responses," in *National Council on Measurement in Education Meeting*, 2011.

[3] 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, October 2009.

[4] J. Bernstein, A. V. Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, p. 355, 2010.

[5] J. Tao, K. Evanini, and X. Wang, "The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 294–299.

[6] L. Chen, "Applying feature bagging for more accurate and robust automated speaking assessment," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 473–477.

[7] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 1188–1196.

[8] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children English language learners," in *Proc. of INTERSPEECH*, 2014, pp. 1468–1472.

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

[10] R. C. van Dalen, K. M. Knill, and M. J. F. Gales, "Automatically grading learners' English using a Gaussian process," in *SLaTE Workshop*, 2015.

[11] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *Proc. of the IEEE ICASSP*, 2016, pp. 6140–6144.

[12] J. Tao, "Acoustic model adaptation for automatic speech recognition and animal vocalization classification," Ph.D. dissertation, Marquette University, 2009.

[13] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Černockỳ, "ivector-based discriminative adaptation for automatic speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 152–157.

[14] 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 the IEEE ICASSP*. IEEE, 2014, pp. 6334–6338.

[15] S. Garimella, A. Mandal, N. Strom, B. Hoffmeister, S. Matsoukas, and S. H. K. Parthasarathi, "Robust i-vector based adaptation of dnn acoustic model for speech recognition," in *Proceedings of Interspeech*, 2015, pp. 2877–2881.

[16] S. Xie and L. Chen, "Evaluating unsupervised language model adaptation methods for speaking assessment," *NAACL/HLT*, pp. 288–292, 2013.

[17] T. Miller, "Essay assessment with latent semantic analysis," *Journal of Educational Computing Research*, vol. 29, no. 4, pp. 495–512, 2003.

[18] S. Xie, K. Evanini, and K. Zechner, "Exploring content features for automated speech scoring," in *Proceedings of NAACL*. Association for Computational Linguistics, 2012, pp. 103–111.

[19] D. Higgins and J. Burstein, "Sentence similarity measures for essay coherence," in *Proceedings of the 7th International Workshop on Computational Semantics*, 2007, pp. 1–12.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Workshop at ICLR*, 2013.

[21] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. of INTERSPEECH*, 2015.

[22] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. of the IEEE ICASSP*, 2014.

[23] M. Gibson, "Minimum bayes risk acoustic model estimation and adaptation," Ph.D. dissertation, University of Sheffield, 2008.

[24] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[25] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of ICSLP*, vol. 2, Denver, USA, 2002, pp. 901–904.

[26] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and audio processing, ieee transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[27] M. J. Gales and P. C. Woodland, "Mean and variance adaptation within the mllr framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[28] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[29] K. Zechner, "What did they actually say? agreement and disagreement among transcribers of non-native spontaneous speech responses in an english proficiency test," in *Proc. of the ISCA SLaTE Workshop*, 2009, pp. 25–28.

[30] A. Loukina, K. Zechner, L. Chen, and M. Heilman, "Feature selection for automated speech scoring," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2015, pp. 12–19.