EXPLORING DEEP LEARNING ARCHITECTURES FOR AUTOMATICALLY GRADING NON-NATIVE SPONTANEOUS SPEECH

Jidong Tao, Shabnam Ghaffarzadegan, Lei Chen, Klaus Zechner*

Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA
* Center for Robust Speech Systems, University of Texas at Dallas, TX 75080, USA
{jtao,lchen,kzechner}@ets.org, shabnam.ghaffarzadegan@utdallas.edu

ABSTRACT

We investigate two deep learning architectures reported to have superior performance in ASR over the conventional GMM system, with respect to automatic speech scoring. We use an approximately 800-hour large-vocabulary non-native spontaneous English corpus to build three ASR systems. One system is in GMM, and two are in deep learning architectures - namely, DNN and Tandem with bottleneck features. The evaluation results show that the both deep learning systems significantly outperform the GMM ASR. These ASR systems are used as the front-end in building an automated speech scoring system. To examine the effectiveness of the deep learning ASR systems for automated scoring, another non-native spontaneous speech corpus is used to train and evaluate the scoring models. Using deep learning architectures, ASR accuracies drop significantly on the scoring corpus, whereas the performance of the scoring systems get closer to human raters, and consistently better than the GMM one. Compared to the DNN ASR, the Tandem performs slightly better on the scoring speech while it is a little less accurate on the ASR evaluation dataset. Furthermore, given the results of the improved scoring performance while using fewer scoring features, the Tandem system shows more robustness for scoring task than the DNN one.

Index Terms— automatic speech scoring, non-native spontaneous speech, automatic speech recognition, deep neural network, bottleneck features

1. INTRODUCTION

For the past two decades, a large number of studies have been conducted using automatic speech recognition (ASR) technology in the assessment of speech, such as computer aided pronunciation training (CAPT) and automated speech scoring (see [1] for a comprehensive review). In automated speech assessment systems, such as those exemplified in [2, 3], different speech features are computed using various methods including signal processing, prosodic analysis, and natural language processing (NLP). The extracted features are fed into a statistical model to automatically predict human speaking proficiency levels.

As systematically investigated in [4], an ASR module used inside an automated speech scoring system plays a very important role in achieving high scoring accuracy. As shown in Section 2, there are emerging efforts in applying the new generation of deep neural network (DNN) based ASR systems to speech assessment tasks. In this paper, we will report our work on (a) using a very large non-native English corpus to train deep learning based ASR systems, (b) comparing the ASR performance of these systems to a conventional Gaussian mixture model (GMM) based one, (c) applying the ASR systems to the automated speech scoring task, and finally (d) comparing the scoring performance. To our knowledge, this is the first ASR study to use approximately 800 hours of non-native spontaneous speech covering over 100 first languages (L1s) across 8700 speakers from about 150 countries around the world; and no work so far has investigated a comparison between deep learning architectures for automated speech scoring.

The paper is organized as follows: Section 2 briefly reviews previous studies on deep learning based ASR and its increasing use in the task of automated speech scoring; Section 3 describes the three types of ASR systems, i.e., GMM, DNN, and Tandem using acoustic features learned from a bottleneck (BN) network; Section 4 describes our speech scoring model setup, including the features and the machine learning prediction models; Section 5 reports our experiments; and finally Section 6 makes conclusions.

2. PREVIOUS RESEARCH

From ASR outputs and prosodic analyses, a set of features suggested widely in English Language Learner (ELL) studies can be automatically computed, such as features measuring speaking fluency, intonation, vocabulary, and so on. One example of a task in speech assessment focuses on measuring pronunciation performance, in which Goodness of Pronunciation (GOP) [5] is a predominant approach to calculate the pronunciation features based on posterior probabilities from an acoustic model (AM) in ASR. A working example of a rich set of speech features can be found in [2].

Non-native speech presents challenges for automatic recognition over and above those presented by native speech. Non-native speech contains more diverse allophones, a broader range of accent, a higher possibility of hesitations, filler words, partial words, or even words randomly invented by speakers, etc. From the ASR standpoint, in order to build an accurate AM for non-native speech, a very large size training corpus with hundreds of hours of speech files is typically required. However, due to the lack of such large size training corpora, most previous research has focused on using various model adaptation approaches to adapt existing models trained on native speech data towards a non-native ASR system [6].

For decades, GMMs have been regarded as the most powerful model for building AMs for ASR systems. Very recently, artificial neural networks with multiple layers have replaced GMMs for building AMs. [7] provides a systematic overview of this fundamental change in the ASR research field. With DNN's widespread successes in ASR, it has also been recently applied to the task of speech assessment. In [8], deep belief network (DBN) was applied to ASR for improving the performance of a CAPT system. [9, 10] first applied DNN in evaluating English learners' pronunciation in

a Computer-Aided Language Learning (CALL) scenario. Multi-layer, stacked Restricted Boltzman Machines (RBMs) were trained as AMs for computing three GOP-style pronunciation features. It was found that DNN AMs improved pronunciation evaluation performance over their GMM counterparts. [11, 12] investigated the use of context-dependent DNN hidden Markov models (CD-DNN-HMM), to improve ASR and obtained more accurate automatic assessment of child English learners. Their DNN-HMM ASR outperformed the GMM-HMM one significantly. Using content and manner-of-speaking features derived from DNN-ASR outputs, machine scoring was improved to the level of human raters. [13] reported training a Tandem GMM-HMM ASR using bottleneck features on 100-hours of non-native English data from the AMI meeting. They obtained a 37.6% word error rate on the BULATS (Business Language Testing Service) corpus of learners' speech made available by Cambridge English.

In previous efforts in applying DNN-HMM ASR systems to speech assessment tasks, researchers mostly focused on one single type of network architecture, namely, DBN. However, very recently, more network architectures have been proposed, and some have already shown promising improvements in speech recognition accuracy [14, 15]. It is worthwhile to evaluate the potential benefits of these different types of neural networks for the speech scoring task.

3. ASR SYSTEMS

In our study, three types of ASR systems with different AMs are built using Kaldi [19], a state-of-the-art open-source ASR toolkit. While the AMs for these ASR systems are different, they all share the same tri-gram language model (LM) for this study. The details of these systems are described below. Note that all of these ASR systems are gender-independent.

GMM system: In the conventional GMM-HMM ASR, 3-state left-to-right context-dependent HMMs with 8 Gaussian mixtures per state are used to model 39 phones and the variants in US-English, plus 1 silence phone, and 2 noise phones. 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. The mean and variance normalized MFCC features are spliced in time taking a context size of 9 frames (4 on each side of the current frame), and are further de-correlated and reduced dimensionality to 40 by applying linear discriminant analysis (LDA) [20]. The classes for the LDA estimation are the triphone states. The resulting features are further de-correlated using maximum likelihood linear transform (MLLT) [21]. The speaker normalization is followed by applying feature-space maximum likelihood linear regression (fMLLR) [22]. The final GMM system in our baseline has 40×41 fMLLR parameters with the speaker adaptive training (SAT) [23] on top of it.

DNN system: In this system, a 5-layer DNN with p-norm (p=2) nonlinearity is trained using layer-wise supervised backpropagation training [24]. A normalization component right after each hidden layer is applied to keep the training stable and prevent the neurons from becoming over-saturated. The same acoustic features as in the final baseline GMM system were used with the same +/- 4 context size of frames. The network training randomly initializes with one hidden layer, trains it shortly, then removes the layer of weights that go to the softmax layer, adds a new hidden layer and two sets of randomly initialized weights, and trains again. The process is iterated until the desired number of layers are produced. Senone state-level posterior probabilities from the output of the DNN are further converted into likelihoods by dividing by the prior of the states, and are fed to the HMM framework as a replacement for the GMM likelihoods.

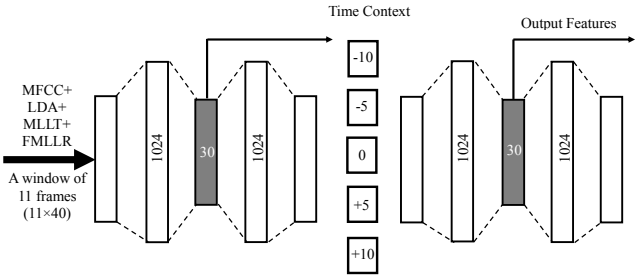

Fig. 1. Two layers stacked bottleneck network [25].

Tandem system: Another class of hybrid ASR systems are Tandem, in which a neural network is used to extract features for GMM-HMM models. The main advantage of using Tandem features is that one can re-use the existing ASR engines by simply replacing the acoustic feature extractor with the Tandem one. As established in the literature [26, 25, 27], Tandem features can be extracted from the output of the neural network or the bottleneck (BN) layer. Bottleneck features (BNF) do not require de-correlation and dimension reduction and gain further improvement compared to traditional Tandem features [26]. Figure 1 shows a BN network used in this study which is a stack of two BN systems [25]. The second BN layer in this architecture can make a better decision based on a longer temporal context. In our case, each BN network is a 5 layer neural network with the BN size of 30 neurons, and hidden layer size of 1024 neurons. The first BN is trained on 11 frames of mean and variance normalized MFCC static, delta and acceleration coefficients. The output of this network from the BN layer is mean and variance normalized and concatenated with time offsets of -10, -5, 0, 5, 10 to prepare the input for the second BN network. The output of the second BN network is used as the final feature for GMM-HMM models. In the network structures for BN training, *sigmoid* is used as the activation function. The cross-entropy criterion has been employed for network training. To achieve the best performance, sequence-discriminative training using maximum mutual information (MMI) is applied on top of the cross-entropy method [28].

4. SCORING SYSTEM

SpeechRater^{SM}, an automated scoring engine for assessing non-native English proficiency [2], is used to extract scoring features and predict a numerical score for spoken responses. The features are related to several aspects of the speaking construct[1], which include *fluency*, *rhythm*, *intonation* & *stress*, *pronunciation*, *grammar*, and *vocabulary* use. A group of these features measuring generic speaking skills was extracted for scoring spontaneous non-native speech. Table 1 provides a concise synopsis of these features.

Based on the available speech features reported in Table 1, automatic scoring feature selection based on LASSO regression [29] is used to obtain a much smaller input feature set (# features are about 33) for building a linear regression model for score prediction. The number of LASSO regression selected features are 26, 33, and 28 for the three different scoring models with GMM, DNN, and Tandem as the front-end respectively. Note that linear regression (LR) is

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

Category	Quantity	Example Features
Fluency	18	Features based on the number of words per second, number of words per chunk, number of silences, average duration of silences, frequency of long pauses ($\geq$ 0.5 sec.), number of filled pauses (*uh* and *um*) [2]. Frequency of between-clause silences and edit disfluencies compared to within-clause silences and edit disfluencies [16].
Rhythm, Intonation & Stress	12	Features based on the distribution of prosodic events (promincences and boundary tones) in an utterance as detected by a statistical classifier (overall percentages of prosodic events, mean distance between events, mean deviation of distance between events) [2] as well as features based on the distribution of vowel, consonant, and syllable durations (overall percentages, standard deviation, and Pairwise Variability Index) [17].
Pronunciation	9	Acoustic model likelihood scores, generated during forced alignment with a native speaker acoustic model, the average word-level confidence score of ASR and the average difference between the vowel durations in the utterance and vowel-specific means based on a corpus of native speech [18]
Grammar	13	Similarity scores of the grammar of the response in ASR with respect to reference response.
Vocabulary Use	13	Features about how diverse and sophisticated the vocabulary based on the ASR output.

Table 1. Descriptions of SpeechRater^{SM} features for automated scoring

used (instead of other more powerful machine learning algorithms) to obtain a more interpretable model.

5. EXPERIMENTAL RESULTS

5.1. Non-native Spontaneous English Corpora

We use TOEFL Internet-based test® (iBT) data in our ASR experiments and TOEFL Practice Online® (TPO) data in our automated scoring experiments. 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. TPO is a practice test provided by TOEFL for test-takers to prepare for the iBT exams. Although both corpora are in the same TOEFL® family, the test administration years and contexts are different. In addition, due to the fact that TPO is not a formal test, English learners may be less motivated for an online practice test. Moreover, they use their own audio input devices and may experience different background noise. Therefore, the TPO corpus used for scoring purposes can be treated as the out-of-domain data to the iBT corpus for ASR training usage.

Table 2 provides more details about these data sets. Regarding the ASR experiments using the iBT corpus, we use a typical data-splitting setup, i.e., asr-train, asr-dev, and asr-eval partitions with manual orthographic transcriptions for ASR training, development and evaluation. Note that there is no speaker overlap among the three sets. The ASR training partition 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. In the automated scoring experiments, 1000 TPO test takers were chosen to create the scoring corpus (each speaker responded to 6 test items to produce 6 spoken responses) , and two thirds of the 6000 total responses were allocated to the scoring model training (sm-train) partition, and one third was allocated to the scoring model evaluation (sm-eval). All responses used for scoring were double scored by experienced human raters following the 4-point scale scoring rubric designed for scoring the TOEFL test. 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 response item level and the speaker level inter-rater agreement are $r_{item} = \kappa_{item} = 0.59$, and $r_{spkr} = \kappa_{spkr} = 0.88$ respectively.

Corpus	Partition	#Items	#Speakers	Dur (hrs)
TOEFL iBT	asr-train	52200	8700	819
	asr-dev	600	100	9.4
	asr-eval	600	100	9.4
TPO	sm-train	4002	667	57.6
	sm-eval	1998	333	29

Table 2. Non-native spontaneous English corpora. #Items: number of responses per set; #Speakers: Number of speakers per set; Dur (hrs): Duration of each set in hours.

5.2. ASR performance

Table 3 shows the results of different ASR systems for the non-native spontaneous speech recognition tasks, namely the word error rate (WER) for asr-eval (WER_{ae}), sm-train (WER_{st}), and sm-eval (WER_{se}) partitions, respectively. A shared LM by the three ASR systems are trained on the asr-train partition. The ASR performance on the asr-eval data using the best conventional GMM system, which is fMLLR with SAT, has a WER of 29.43%. Using DNN instead of GMM in the HMM frame-work achieves a 22.76% WER, which is about 23% relative WER reduction over the GMM system. The Tandem system achieves a 23.07% WER, which is about 22% in relative WER reduction over GMM. The significant WER reduction on both deep learning based ASR systems shows that a high performance gain in non-native spontaneous ASR is achievable. The ASR performance on the two scoring model partitions in Table 3 show that all three ASR systems have absolute WER increments in the close range of 12-13%. One possible reason for the substantial ASR per-

System	WER_{ae}	WER_{st}	WER_{se}
GMM	29.43	43.35	42.45
DNN	22.76	35.41	35.11
Tandem	23.07	35.17	35.07

Table 3. WER for 3 data partitions using 3 different ASR systems.

System	r_{item}	κ_{item}	r_{spk}	κ_{spk}
GMM	0.52	0.48	0.74	0.73
DNN	0.55	0.52	0.76	0.74
Tandem	0.58	0.53	0.78	0.78

Table 4. Pearson correlation (r) and quadratic weighted kappa (κ) between SpeechRater^{SM} and human raters' scores in item and speaker level across 3 ASR systems.

formance drop is the acoustic condition and quality of the TPO data, and another is that the spoken responses are out-of-domain. The calculated LM perplexities are 69.6, 132.0, and 130.1 for the asr-eval, sm-train, and sm-eval partitions respectively. The perplexity results indicate 1) that the TPO scoring corpus in context is out-of-domain to the iBT ASR training corpus; 2) a ballpark lower bound for deep learning based ASR systems in operational use without the pre-knowledge of data.

5.3. Scoring performance

Table 4 reports our machine scoring experiment using the three trained ASR systems. In order to simulate in a real world scenario without any pre-knowledge of whether the spoken responses in the scoring data are from the same domain as the ASR training set, all three ASR systems are built using the ASR training partition only, without additional tuning or optimization towards the scoring data. Both deep learning based scoring systems perform considerably better than GMM, which is reflected by both the item and speaker level correlations' increment. In automated speech scoring, quadratic weighted kappa (κ) normally has a lower value than Pearson correlations (r). The difference between the two correlations can indicate how efficiently and robustly the scoring system can predict consistent scores in the both float point and integer. Between two deep learning based scoring systems, the Tandem system shows more robustness and efficiency with the highest correlations to human rater in all cases, while the ASR performance is similar to DNN. The results indicate that deep learning approaches in the acoustic feature domain have a bigger impact on automated scoring systems than those in the model domain.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we explored two deep learning architectures to improve an automated scoring system for non-native spontaneous speech using a very large corpus for ASR training. Both deep learning based ASR systems substantially reduce WER compared to the best conventional GMM system. In a similar manner, using the deep learning ASR systems as the front-end of an automated scoring system shows a considerable gain over the GMM based scoring system. Compared to the ASR performance on the in-domain data, the recognition accuracy of each system drops significantly on the out-of-domain data. Between the two deep learning based ASR systems, the Tandem performs slightly better on the out-of-domain speech while it is a little less accurate on the in-domain data. Furthermore, the Tandem system shows more robustness in scoring, given the experimental results of the consistently improved scoring performance while using fewer of the selected scoring features.

The conclusions in this paper were drawn using one non-native spontaneous speech corpus for building ASR systems, and another corpus for the scoring purpose. That the scoring corpus was out-of-domain to the ASR training data caused the ASR performance drop. It will be worthwhile to examine both the ASR and scoring performance on the in-domain corpus in order to better understand how the impact of deep learning based speech assessment differs between the in-domain and out-of-domain data. In the case of a lack of a human transcribed scoring corpus, a possible ASR performance improvement on out-of-domain data can be archived using LM adaptation approaches. The two deep learning ASR systems were trained by following the original publications from two different groups of authors; the two systems differed in complexities such as the number of context frames and the number of network layers, etc. Future work will compare the complexities of the different deep learning architectures. Our scoring results showed a general direction of the deep learning approaches in the field of speech assessment; future studies will focus on the scoring features used in the different scoring models across different test takers' scores and L1s in order to enhance the interpretation of deep learning methods in speech scoring.

7. REFERENCES

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

[2] 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.

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

[4] 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.

[5] S. M. Witt, *Use of Speech Recognition in Computer-assisted Language Learning*, Ph.D. thesis, University of Cambridge, 1999.

[6] L. M. Tomokiyo, *Recognizing non-native speech: characterizing and adapting to non-native usage in LVCSR*, Ph.D. thesis, Carnegie Mellon University Pittsburgh, PA, 2001.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[8] X. Qian, H. Meng, and F. K. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," in *Proc. of Interspeech*, 2012.

[9] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL).," in *Proc. of INTERSPEECH*, 2013, pp. 1886–1890.

[10] 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.

[11] 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.

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

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

[14] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. Mohamed, G. E. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[15] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. of the IEEE ICASSP*, 2012, pp. 4277–4280.

[16] L. Chen and S. Y. Yoon, "Application of structural events detected on asr outputs for automated speaking assessment," in *Proc. of INTERSPEECH*, 2012.

[17] L. Chen and K. Zechner, "Applying rhythm features to automatically assess non-native speech," in *Proc. of INTERSPEECH*, 2011, pp. 1861–1864.

[18] L. Chen, K. Zechner, and X Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *NAACL-HLT*, 2009.

[19] 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. Vesely, "The Kaldi speech recognition toolkit," in *Proc. of the IEEE ASRU*, 2011.

[20] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. of the IEEE ICASSP*, 1992, pp. 13–16.

[21] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, pp. 75–98, 1998.

[22] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. of the IEEE ICASSP*, 1998, pp. 661–664.

[23] S. Matsoukas, R. Schwartz, H. Jin, and L. Nguyen, "Practical implementations of speaker-adaptive training," in *DARPA Speech Recognition Workshop*, 1997.

[24] 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.

[25] K. Veselý, M. Karafiát, and F. Grézl, "Convolutive bottleneck network features for LVCSR," in *Proc. of the IEEE ASRU*, 2011, pp. 42–47.

[26] F. Grézl, M. Karafiát, S. Kontar, and J. Cernocký, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. of the IEEE ICASSP*, 2007, pp. 757–760.

[27] F. Grézl and P. Fousek, "Optimizing bottle-neck features for lvcsr," in *Proc. of the IEEE ICASSP*, 2008, pp. 4729–4732.

[28] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. of INTERSPEECH*, 2013, pp. 2345–2349.

[29] 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.