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

Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment

Yao Qian, Xinhao Wang, Keelan Evanini, David Suendermann-Oeft

Educational Testing Service Research, USA
{yqian, xwang002, kevanini, suendermann-oeft}@ets.org

Abstract

Automated assessment of language proficiency of a test taker's spoken response regarding its content, vocabulary, grammar and context depends largely upon how well the input speech can be recognized. While state-of-the-art, deep neural net based acoustic models have significantly improved the recognition performance of native speaker's speech, good recognition is still challenging when the input speech consists of non-native spontaneous utterances. In this paper, we investigate how to train a DNN based ASR with a fairly large non-native English corpus and make it self-adaptive to a test speaker and a new task, namely a simulated conversation, which is different from them monologic speech in the training data. Automated assessment of language proficiency is evaluated according to both task completion (TC) and pragmatic competence (PC) rubrics. Experimental results show that self-adaptive DNNs trained with *i*-vectors can reduce absolute word error rate by 11.7% and deliver more accurate recognized word sequences for language proficiency assessment. Also, the recognition accuracy gain translates into a gain of automatic assessment performance on the test data. The correlations between automated scoring and expert scoring could be increased by 0.07 (TC) and 0.15 (PC), respectively.

Index Terms: speech recognition, non-native spontaneous speech, automated speech scoring, DNN, *i*-vectors

1. Introduction

In recent years, deep learning in combination with large databases have significantly improved performance and system robustness of speech recognition, dialogue management, language understanding and machine translation [1-3]. This has also resulted in an increased interest in conversation-based computer-assisted language learning (CALL), in which formative assessment of spoken language proficiency is a core component.

Traditionally, automatic assessment of spoken language is performed on restricted speech, e.g., reading a sentence out loudly. This is due to the difficulty in obtaining accurate ASR output to be used for scoring the responses, especially for non-native speech that may contain pronunciation errors, high amounts of disfluencies, ungrammatical phrases, etc. [4] Recently, with the significantly improved discrimination of acoustic modeling by deep learning in ASR, several automated assessment systems were developed to score spontaneous speech [5-8]. The comparison between DNN-HMM and GMM-HMM for acoustic modeling shows that DNN-HMM can significantly increase speech recognition performance of test takers' responses, and improve the quality of the extracted features for automatic assessment of spoken language proficiency, and consequently achieve a performance close to that of human scorers.

However, most systems are task-dependent, i.e., building acoustic models and scoring models for different datasets is usually data- and time-consuming. In real applications, the pilot data used for building the original ASR is often not enough to produce output that is accurate enough to be employed for assessing spontaneous responses. [8] directly applies ASR to multiple tasks and domains without adaptation, counting on the inherent relative robustness of DNN over GMM for acoustic modeling. Despite the ASR accuracy improvement, the performance of automated assessment is still suffering. The idea of linear transformation, which is generally used to GMM adaptation, is not applicable to DNN, since it would require a nonlinear transformation for layer connections and discriminative training with back-propagation (BP) rather than maximum likelihood (ML) training with expectation–maximization (EM). Currently, the effective adaptation of DNN-based acoustic models for ASR is an active research area [13]. To our knowledge, there are few research studies investigating the adaptation of DNN-HMM with limited data for the purpose of assessing non-native spontaneous speech.

In this paper, we explore the adaptation of ASR built on a very large non-native English corpus to the assessment of language proficiency for non-native spontaneous speech in a simulated conversion. A comparison of adapting acoustic models between GMM-HMM and DNN-HMM and interpolating language models from monolog tasks to simulated dialog tasks is drawn. Features extracted from the ASR output with different adaptation strategies are also surveyed for their correlation to human scores and their contribution to the automated scoring models.

2. Data and Task

Two non-native spontaneous English corpora are used in this study. The first is drawn from a large-scale standardized spoken language proficiency test which measures a non-native speaker's ability to use and understand English at the university level. The speaking tasks in this test elicit monologs of 45 or 60 seconds in duration; example tasks include expressing an opinion on a familiar topic or summarizing information presented in a lecture. Human experts were recruited to rate the responses using holistic rubrics on a 1-4 scale that cover the following three main aspects of speaking proficiency: delivery, language usage and topic development. This corpus is hereafter referred to as the monolog corpus.

The second corpus is drawn from a pilot administration of an assessment of non-native English speaking proficiency for academic purposes. This assessment contained a task type in which the test taker is presented with a set of stimulus materials, such as a course schedule, an advertisement for a job on

http://dx.doi.org/10.21437/Interspeech.2016-291

campus, etc., and is then presented with a series of spoken prompts from a computer-based interlocutor in the form of a simulated dialog. After each response from the test taker, the subsequent prompt from the computer-based interlocutor is played, until the final prompt has been reached. Expert human raters provided proficiency ratings for an entire simulated dialog, i.e., a rater listened to all of the responses provided by a test taker in the conversation and then provided a single score for the entire conversation. Ratings were given on a scale of 1-5 for the following two dimensions of speaking proficiency: Task Completion (TC), which demonstrates an ability to complete the communicative demands of the task by providing the required content from the stimulus materials and using correct grammar and vocabulary, and Pragmatic Competence (PC), which demonstrates an ability to use language that is appropriate to the context (situation and role). The corpus and scoring rubrics are described in further detail in [4]. This corpus is hereafter referred to as the dialog corpus.

The monolog corpus contains over 800 hours of non-native spontaneous speech covering over 100 L1s (native languages) across 8,900 speakers. The acoustic and language models of our ASR system were trained on this corpus. Table 1 presents the number of speakers, number of responses and duration of speech for three partitions that were used for building the ASR system: training (AsrTrain), development (AsrDev), and evaluation (AsrEval); there is no speaker overlap across the three partitions.

Table 1. *Number of speakers, responses and duration of speech for each data partition in the monolog corpus.*

	AsrTrain	AsrDev	AsrEval
# Speakers	8,700	100	100
# Responses	52,200	600	600
# Hours	819	9.4	9.4

The dialog corpus consists of 1,922 test takers representing 51 L1s. The 1,922 conversations (10,276 responses) from the simulated dialog task are divided into the following four sets (with no speaker overlap) for the current study: ASR adaptation (AsrAdapt), ASR evaluation (AsrEval), scoring model training (SmTrain) and scoring model evaluation (SmEval). The corresponding number of speakers, responses and hours are presented in Table 2.

Table 2. *Number of speakers, responses and duration of speech for each data partition in the dialog corpus.*

	AsrAdapt	AsrEval	SmTrain	SmEval
#Speakers	584	201	860	277
#Responses	3,155	1,072	4,585	1,464
#Hours	26	9	38	12

3. Speaker Adaptation of DNN

Although DNN based ASR systems show superior generalization capability than those based on GMM, they still suffer from a mismatch between training models and testing speakers' data, which can be caused by variation of acoustic environment, speakers and task domain. Our system built on the monolog corpus has a large degradation of recognition accuracy when ported to unseen speakers and domains [8]. Our conjecture is that speaker mismatch here is not only caused by the variation of speaker characteristics, e.g, vocal tract length and speaking style, but also by the variation of L1, due to the mismatch between the nature of the recordings in the monolog and dialog corpora.

There are a considerable number of approaches to investigate the feature transformation of test speakers towards trained models [9], models (or certain layers of models) re-update towards testing speakers [10], or training with additional speaker information [11-13]. DNNs, unlike GMMs, do not have a straightforward way to adapt models due to significantly more parameters of deep hidden layers and distributed training. Since there are no overlapping speakers between the two tasks in this study, the unsupervised approaches of feature-space adaptation and adding speaker information are explored.

3.1. Adapting DNNs with fMLLR

Feature-space maximum likelihood linear regression (fMLLR) [14], is an affine feature transform of the form

$$\hat{x} = Ax + b \qquad (1)$$

where a d by d transformation matrix A and a d-dimensional bias vector b are estimated by aiming to maximize the likelihood of adaptation data. fMLLR transformed features are used as the input to the DNN. It is regarded as feature normalization rather than the original concept of fMLLR adaptation of GMM, in which the transform needs to be estimated in the maximum likelihood sense with the EM algorithm. We use unsupervised 2-pass estimation in the decoding procedure. The first pass decoding is performed with GMM-HMM models. The lattice outputs by the first pass are used to compute fMLLR transforms on each speaker, and the second pass decoding is performed using a DNN trained on fMLLR transformed acoustic features, which are also produced by GMM-HMM.

3.2. Adapting DNNs with *i*-vectors

Based upon factor analysis, an *i*-vector is a compact representation of a speech utterance in a low-dimensional subspace [15, 16]. The i-vector approach has become the state-of-the-art in the speaker recognition field. Given a GMM, the corresponding mean super-vector M can be approximated by:

$$M = m + T\omega \qquad (2)$$

where m represents a speaker- and channel-independent supervector, which can be estimated by a UBM; T is a low-rank matrix representing the total variability of speaker and channel across the collected data. ω is the *i*-vector, a low-dimensional vector with a normally distributed prior $N(0; I)$, estimated by the EM algorithm over the training corpus. We extract an *i*-vector for each speaker by using a pre-trained T matrix and append this feature vector with the acoustic feature vector as input to the DNN training and later the recognition. Length normalization [17] is applied to the *i*-vector to sharpen its Gaussian distribution and match the distributions of training and testing. In addition, to avoid over-fitting of DNN training with *i*-vectors, the weight decay parameter of *l2* regularization is adjusted.

Although both fMLLR transformed features and i-vector appended features are used to adapt the DNN, they seem to have different physical meanings. fMLLR is a technique to remove speaker variability and the corresponding transformed features are supposed to no longer contain information related to the speaker. DNN trained based on fMLLR transformed features should be regarded as speaker independent (SI) models. However, i-vectors are to capture all the speaker-specific information in a speaker sensitive subspace. A DNN trained on features with appended i-vectors is to learn speaker-specific information for improving its robustness to unseen speakers, whose i-vectors can be extracted by the T matrix.

4. Language Proficiency Assessment

Over 100 features that cover a range of linguistic characteristics of the spoken response, such as fluency, intonation, stress, rhythm, pronunciation, vocabulary, and grammar, were extracted for assessing a language learner's English speaking proficiency [18]. The feature extraction is performed by using a two-pass approach that first conducts ASR on the spoken response using acoustic and language models trained from non-native spontaneous speech and then conducts forced alignment of the spoken response to the ASR output using an acoustic model trained on native speech. The non-native ASR is mainly used to extract the features that address the appropriateness of content, vocabulary, grammar and context usage, while native ASR is employed to extract the features used to evaluate pronunciation, fluency, and intonation against nativeness. In this study, we explore the impact of improved non-native ASR on the assessment of language proficiency. The following three features, which are extracted from non-native ASR and resulted in the highest correlations with human scores, are investigated in detail by adapting the non-native ASR from monolog to dialog tasks.

LSA: Latent semantic analysis is used to extract contextual meaning of words by statistical computations for a large corpus of text. LSA can estimate the quality and quantity of knowledge contained in an essay and then assess the learnability of passages by individual students [24]. Here a test taker's spoken response is graded by the similarity between recognized word and word sequences from test taker and the training set, projected in a reduced space.

CVA: Content Vector Analysis is another approach to estimating the appropriateness of spoken content [25]. CVA is similar to LSA in that it uses cosine similarity measures, but no SVD is employed in CVA. In addition, CVA differs from LSA in that CVA divides responses into groups of human scores [4].

CS: Confidence score is a measure of confidence for the recognized word or other units from ASR. Generally, CS is the posterior probability of the ASR output and used as an invaluable source of information when incorporating the ASR output into other systems. We use the average of word confidence score per response as feature.

A statistical model is used to predict the final proficiency score from all features provided.

5. Experiments and Results

ASR systems are constructed by using the tools from Kaldi [19] and CNTK [27].

5.1. Experimental setup

GMM-HMM and DNN-HMM are trained by using the data set of AsrTrain from monolog corpus.

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

The features used to train DNN are MFCC features with the same dimensions as those used in GMM-HMM. The input features stacked over a 15 frame window (7 frames to either side of the center frame for which the prediction is made) are used as the input layer of DNN. The output layer of DNN has 4057 nodes, i.e. senones of the HMM obtained by decision-tree based clustering. The input and output feature pairs are obtained by frame alignment of the senones with GMM-HMM. The DNN has 7 hidden layers, each layer with 1024 nodes. The sigmoid activation function is used for all hidden layers. All the DNN parameters are initialized by layer-wise BP pre-training [20], then trained by optimizing the cross-entropy function through back-propagation, and finally refined by Sequence-discriminative training using state-level minimum Bayes risk (sMBR).

Speaker adaptive training is performed on GMM-HMM. The 40×41 fMLLR transform is estimated for each speaker upon LDA and MLLT normalized features by the EM algorithm with GMM-HMM. fMLLR transformed features are fed into the training of DNN-HMM. AsrTrain of the monolog corpus is also used to train hyper-parameters: GMM-UBM and T-matrix for i-vector extraction. A 100 dimensional i-vector is extracted to be stacked with original DNN input features per frame to train DNN-HMM.

The CMU pronunciation dictionary [21] is used to build a grapheme-to-phoneme (G2P) converter by data-driven joint-sequence models [22]. After text normalization for transcriptions, we use G2P to automatically generate pronunciations for each word in the transcription and combine them with the CMU dictionary to create a new pronunciation dictionary. The vocabulary size of AsrTrain in the monolog corpus is 23,144. There are 582 unseen words in the AsrAdapt set of dialog. The same G2P converter is used to predict their pronunciations to be added to the pronouncing dictionary.

Two trigram LMs are trained from the transcriptions of the AsrTrain set in the monolog corpus and the AsrAdapt set in the dialog corpus by the IRSTLM toolkit [23], separately. Linear interpolation is used to combine these two LMs. The interpolation weight is optimized by minimizing the WER on the AsrEval set in the dialog corpus. The interpolated (combined) LM is finally represented as a finite state transducer (FST) for weighted FSTs (WFSTs) based decoding.

The native ASR used for extracting nativeness related features, is built by data from the Wall Street Journal CSR Corpus, which contains about 36k utterances recorded by 284 native American English speakers. The scoring model is built on the SmTrain set of the dialog corpus by using a random forest regressor [26], which achieves the best performance over other regressors, e.g., AdaBoost and DecsionTree. The scores for TC and PC are mostly rated by two experts except that the third opinion is given when the scores from those two experts differ by more than one. We use the average scores from experts as the ground truth or reference score to build scoring model.

5.2. ASR results and analysis

Performance results of the ASR for using GMM and DNN for acoustic modeling and speaker adaptation with fMLLR and *i*-vector are listed in Table 3. In can be seen that:

1) The performance of ASR with DNN-based AM drops significantly, i.e., WER is increased from 24.2% to 35.0%, in cross-task speech recognition from monolog to dialog. It is even worse (2.9% higher WER) than the performance of GMM-based AM with fMLLR adaptation, which reinforces our motivation to adapt DNNs toward target speakers in the test domain.

2) Both fMLLR and i-vector are effective for speaker adaptation and result in WER reduction. Speaker adaptation with i-vector, however, achieves the best performance, i.e, the absolute WER improves by 4.1% and 3.6%, respectively, when comparing with fMLLR on the AsrEval sets of monolog and dialog. We conjecture that i-vectors extracted from the monolog corpus convey certain phonetic variations brought by the L1s of test takers in addition to speaker characteristics, transmission channel and acoustic environment. It will be tested for L1 identification in our future work. Furthermore, the *i*-vector, due to its properties, can well interpolate the unseen speakers in reduced-dimension subspace and hence the frame posterior probabilities generated by DNNs are well interpolated for decoding.

3) DNN consistently outperforms GMM for acoustic modeling, with or without speaker adaption. DNN AM with adaptation and DNN+i-vector can significantly improve the performance of ASR, i.e., absolute WER reductions of 7.6% and 8.8% for monolog and dialog, compared with GMM AM with adaption, GMM+fMLLR, which is slightly better than the improvement obtained by going from GMM to DNN without adaptation. DNN capability in modeling longer-span, higher dimensional and correlated input features has made the integration i-vector with DNN seamless. On the other hand, incorporate i-vector with GMM AM is nontrivial.

Table 4 shows the effect of LM interpolation for cross-task speech recognition by evaluating the WER of Dialog AsrEval. The combined LM outperforms either LM trained from Monolog AsrTrain or LM trained from Dialog AsrAdpt. The optimized weight for interpolation is 0.9 assigned to LM trained from Dialog.

5.3. Proficiency assessment results and analysis

The results of language proficiency assessment are shown in Table 5. The content related features: LSA, CVA and CS, extracted from different ASRs, i.e., AM trained by GMM and DNN with speaker adaptation of fMLLR and i-vector, are evaluated by Pearson correlation coefficient with reference scores (average expert scores). It shows that the higher recognition accuracy achieved by ASR, the higher correlations with the reference (expert) scores. As a reference, the accuracy of content words of nouns, verbs, or adjectives, delivering lexical meaning, rather than indicating a syntactic function, by using different ASRs for Dialog SmEval is listed in Table 6. Table 5 also shows that CVA features outperform LSA features for the TC scores, but the opposite phenomena is observed for the PC scores. We imagine that LSA features capture the long contextual information by using frequency of word sequence and is supposed to be more related to the scoring rubrics of PC scores. By using adapted DNN instead of adapted GMM, the performance of the automated scoring model, which employs more than100 features including LSA, CVA and CS, can be significantly improved from 0.77 to 0.82 for TC and from 0.68 to 0.81 for PC. The improvement for PC scores (approximately 0.13) is much larger than for TC scores. We think that rating PC score is more related to the ability of understanding and participating conversion effectively, so the performance of automated scoring should be more affected by the accuracy of ASR outputs. DNN adaptation with *i*-vector can achieve additional 0.02 and 0.03 improvement of correlation coefficient with reference scores for TC and PC scores, respectively, comparing with DNN adaptation with fMLLR.

Table 3. *WER(%) of GMM and DNN acoustic models with fMLLR and i-Vector adaptations on Monolog and Dialog AsrEval sets.*

	Monolog(AsrEval)	Dialog(AsrEval)
GMM	28.5	42.9
GMM+fMLLR	26.1	32.1
DNN	24.2	35.0
DNN+fMLLR	22.6	26.9
DNN+I-Vector	**18.5**	**23.3**

Table 4. *WER(%) of language models with different interpolation weights on Dialog AsrEval sets.*

Weight	GMM+fMLLR	DNN+I-Vector
1	33.8	24.5
0.9	**32.1**	**23.3**
0.7	32.4	23.5
0.5	32.9	23.9
0 (w/o)	45.2	33.1

Table 5. *Correlations of individual features: LSA, CVA and CS generated by different ASRs, with reference scores (average expert scores), correlations between automated scores from scoring model (SM) and reference scores.*

	Task Completion				Pragmatic Competence			
	LSA	CVA	CS	SM	LSA	CVA	CS	SM
GMM+fMLLR	0.48	0.53	0.42	0.77	0.55	0.54	0.42	0.68
DNN	0.47	0.53	0.41	0.75	0.54	0.52	0.39	0.66
DNN+fMLLR	0.52	0.60	0.52	0.80	0.60	0.58	0.54	0.78
DNN+i-vector	0.54	0.62	0.56	**0.82**	0.62	0.61	0.57	**0.81**

Table 6. *The accuracy (%) of content words by using different ASRs for SmEval set of Dialog corpus.*

	GMM+fMLLR	DNN+fMLLR	DNN+*i*-vector
SmEval	71.3	78.9	81.9

6. Conclusions

In this paper, we propose an i-vector based, self-adaptive DNN AM and show it can improve speech recognition performance, which in turn improves the language proficiency assessment of spoken input of non-native speakers in a simulated conversation task. Significant WER reduction is observed for the adapted ASR with the i-vector based speaker adaptive AM and interpolation of LM. The resultant better recognized word sequence then significantly boosts the performance of automated language proficiency assessment.

7. Acknowledgements

Authors would like to thank Su-Youn Yoon for providing syllabification model to predict syllable boundaries and Sandeep Singh for training CVA models in this study.

8. References

[1] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks. *Science,* vol. 313. no. 5786, pp. 504 - 507, 2006.

[2] Y. Bengio, "Learning deep architectures for AI", *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[3] G. E. Hinton, L. Deng, D. Yu, 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", *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] K. Evanini, S. Singh, A. Loukina, X. Wang and C. M. Lee, "Content-based automated assessment of non-native spoken language proficiency in a simulated conversation", in *NIPS Workshop on Machine Learning for Spoken Language Understanding and Interaction*, 2015.

[5] 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

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

[7] A. Metallinou and J. Cheng, "Using Deep Neural Networks to improve proficiency assessment for children English language learners," in *Proc. of Interspeech,* pp. *1468–1472,* 2014.

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

[9] S.H.K. Parthasarathi, B. Hoffmeister, S. Matsoukas, A. Mandal, N. Strom and S. Garimella, "fMLLR based feature-space speaker adaptation of DNN acoustic models", in *Proc. of Interspeech*, 2015.

[10] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition, " in *Proc. of ICASSP*, pp. 7893–7897, 2013.

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

[12] 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 ICASSP*, pp. 6334-6338, 2014.

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

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

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

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

[17] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proc. InterSpeech*, pp. 249-252, 2011.

[18] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english, " Speech Communication, vol. 51, pp. 883-895, 2009.

[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. ASRU*, 2011.

[20] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription, " in *proc. of IEEE ASRU*, 2011.

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

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

[23] A. Stolcke, SRILM - An Extensible Language Modeling Toolkit, in Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado, September 2002.

[24] T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to Latent Semantic Analysis, Discourse Processes, 25, 259-284, 1998.

[25] S. Xie, K. Evanini and K. Zechner, "Exploring content features for automated speech scoring," in *Proc. of NAACL HLT*, 2012.

[26] https://github.com/EducationalTestingService/skll

[27] D. Yu, A. Eversole, M.L. Seltzer, K. Yao, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, G. Chen, H. Wang, J. Droppo, A. Agarwal, C. Basoglu, M. Padmilac, A. Kamenev, V. Ivanov, S.Cyphers, H. Parthasarathi, B. Mitra, Z. Huang, G. Zweig, C. Rossbach, J. Currey, J. Gao, A. May, B. Peng, A. Stolcke, M. Slaney, X. Huang, "An introduction to computational networks and the computational network toolkit", *Microsoft Technical Report MSR-TR-2014-112*, 2014.