INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Non-Native Children's Automatic Speech Recognition: the INTERSPEECH 2020 Shared Task ALTA Systems

Kate M. Knill, Linlin Wang, Yu Wang, Xixin Wu, Mark J.F. Gales

ALTA Institute, Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK.
{kate.knill,lw519,xw369,yw396,mjfg}@eng.cam.ac.uk

Abstract

Automatic spoken language assessment (SLA) is a challenging problem due to the large variations in learner speech combined with limited resources. These issues are even more problematic when considering children learning a language, with higher levels of acoustic and lexical variability, and of code-switching compared to adult data. This paper describes the ALTA system for the INTERSPEECH 2020 Shared Task on Automatic Speech Recognition for Non-Native Children's Speech. The data for this task consists of examination recordings of Italian school children aged 9-16, ranging in ability from minimal, to basic, to limited but effective command of spoken English. A variety of systems were developed using the limited training data available, 49 hours. State-of-the-art acoustic models and language models were evaluated, including a diversity of lexical representations, handling code-switching and learner pronunciation errors, and grade specific models. The best single system achieved a word error rate (WER) of 16.9% on the evaluation data. By combining multiple diverse systems, including both grade independent and grade specific models, the error rate was reduced to 15.7%. This combined system was the best performing submission for both the closed and open tasks.

Index Terms: speech recognition, children's speech, language learning

1. Introduction

Learners of a language need to be able to have their progress measured to prove their capability to perform a job or study a course, or to move to the next level. The first stage of automatic spoken language assessment (SLA) systems is automatic speech recognition (ASR) to convert the learner's speech into a transcription that can be scored. Standard English ASR systems struggle with understanding learner speech due to mis-pronunciations and non-grammatical speech, which are heavily influenced by the learner's first language(s) and proficiency level. For both adult, e.g. [1, 2, 3, 4], and children's, e.g. [5, 6, 7, 8, 9], SLA a few non-native English ASR systems have been successfully implemented. High performance, however, is still a challenge, with limited labelled training data available. This is particularly true for children's speech where the task difficulty is compounded by higher levels of acoustic and lexical variability, increased code-switching, and a severe lack of even native speech and text training data. To address this the *INTERSPEECH 2020 Shared Task on Automatic Speech Recognition for Non-native Children's Speech* was proposed.

This paper reports on research supported by Cambridge Assessment, University of Cambridge. Thanks to the families of Linlin Wang and Yu Wang for their support and understanding during the evaluation under the COVID-19 lockdown, especially Yvonne and Emily, YVEM.

The organisers distributed recordings of children's English exams collected in the Trentino region of northern Italy [10]. Two sets of manually transcribed audio training data were released; a 9 hour set collected in 2017 and a 40 hour set collected in 2016 and 2018. In addition a 2 hour development set from 2017 was provided. The evaluation was performed on a held out set of 2.3 hours of 2017 data. A set of answers to a written exam by some of the same children was also provided for language model training. English, German and Italian pronunciation lexicons, a Kaldi ASR recipe and a scoring script completed the material distributed for the Shared Task. This formed the basis of the Closed Track evaluation.

The data is taken from recordings of three sets of exams: 9-10 year olds at CEFR [11] level A1; 12-13 year olds at level A2; and 14-16 year olds at level B1. All levels have the same introductory questions such as "What are your hobbies? Why?" but A2 and B1 speakers are expected to answer with more detail. The second part of the tests moves from a simple pizza ordering role play for A1 to more open ended small talk questions at A2 to expanded role-play questions to provide freedom and creativity to the B1 students. These difference in the questions are reflected in the responses. The minimal English A1 speakers produce very short (average 4 word) responses, whereas the basic English A2 and limited but effective use of English B1 speakers talk a lot more, with an average 20 and 28 words, respectively, measured on the development set (Dev). Similarly the vocabulary used grows with level, from 181 unique English words at A1 to 295 at A2 and 466 at B1 in Dev. An interesting feature of the data is that code-switching is quite high, at 4% of words for A2 and B1 speakers and 9% for A1 speakers. Both Italian and German code-switching occurs, a reflection of the linguistic make-up of the Trentino region.

This paper presents the systems developed at the ALTA Institute for the Shared Task Closed Track and lessons learnt in their creation. Section 2 describes the baseline ASR system implemented. Developments to the system are presented in Section 3 followed by the evaluation systems in Section 4 and conclusions.

2. Baseline ASR System

A Kaldi [12] baseline ASR recipe was distributed with the shared task. This system used 9 hours data of the acoustic model (AM) training data, the 2017 data, with additional language model (LM) training data taken from 2016 written exams. The AM configuration specified was a factorised form of time-delay neural networks (TDNNs), TDNN-F [13], trained with lattice-free maximum mutual information (LF MMI) [14]. There were 13 TDNN-F layers of size 1024. The input features were 40-d high-resolution MFCC, and 100-d online extracted i-vectors provide speaker adaptation. 3-way speed perturbation

http://dx.doi.org/10.21437/Interspeech.2020-2154

was applied to augment the AM training data. A four-gram LM was specified.

As the children's speech is spontaneous it contains disfluencies including hesitations and partial words, and non-speech events like laughter. They also sometimes whisper words and code-switch into their native Italian and German. These are all marked in detail in the 2017 data sets. Mis-pronunciations are also marked. As part of the Kaldi recipe a pre-processing step is run on the transcriptions prior to training. The mis-pronunciation and whisper markers are removed and code-switched German and Italian words tagged in the training data are replaced by the name <unk-de>, <unk-it>, respectively. If a sequence of words are tagged then they are replaced by a single word in the transcription e.g.

```
@de(nord italien)      →   <unk-de>
@it(come si dice fa)   →   <unk-it>
```

These words are modelled by single phones in the distributed English pronunciation lexicon as follows:

```
<unk-de>   unk_de_S
<unk-it>   unk_it_S
```

For the LM training the hesitations and non-speech events are also stripped from the transcriptions.

A scoring tool was distributed to enable system development [1]. At scoring any disfluencies, non-speech events and code-switched words are removed from the reference and hypothesis. Note, Italian and German proper nouns remain. The more detailed 2017 transcription resulted in some inconsistencies with the other set. A post-processing step [2] was therefore added to the ASR output prior to scoring: spellings were made consistent (in particular, "favorite"/"favourite") and some miss spellings corrected; foreign words that were not found in the distributed English lexicon were removed; a few split hyphenated words were rejoined. This post-processing reduced the word error rate (WER) by about 1% absolute.

The distributed system only used the initial 9 hours of acoustic training data and yielded a WER of 36.8% on the development data trained on Kaldi v5.5. Using the same recipe, an updated baseline making use of all the available AM training data, 49 hours, and including these manual transcriptions in an updated LM [3]. This yielded an error rate of 22.7%. The breakdown by grade is given in Table 1. Interestingly the A2 speakers have the lowest WER, possibly reflecting the differences in the form of test used at the different grades.

Table 1: *% WER of baseline system on Dev set, trained on 49 hour data set.*

System	% WER			
	A1	A2	B1	Total
Distributed	26.5	20.6	21.6	22.7
Baseline	23.4	17.4	21.8	21.2
+SpecAug	22.0	17.1	20.6	20.2
+RNNLM	21.7	16.6	18.9	19.1
+su-RNNLM	21.5	16.6	18.5	18.8

[1] `TLT2020EvalScriptV2.pl`

[2] The effect of the difference in the guidance was only noted late in the system development so it was not possible to apply the corrections during model training. Initial experiments indicated little difference in system performance.

[3] LM training data: `TLT16W17train` and, after removing `@` words, `TLT1618train.norm.trn`.

Based on the distributed recipe a modified version was built. The primary difference between the systems was that rather than using the English unknown word symbol in the training transcription (<unk>), the closest English word in the vocabulary was selected using alignment with simple initial PLP GMM-HMM models in HTK v3.4.1 [15]. All code-switched words were replaced with an individual word (<unk-de>/<unk-it>) and hesitation tags were mapped to a single word (%HES%) and this was modelled in both the AM and LM. Additionally minor changes were made to the network configuration, with 15 TDNN-F layers and 40-dim log-filterbank features used. Table 1 shows the impact of these changes, reducing the WER by 1.5% absolute. For these forms of low-resource tasks, data augmentation is a commonly used approach. In addition to speed perturbation, SpecAugment [16] was also applied, where time and frequency bands of the spectrogram are randomly masked out in training. The distributed Kaldi implementation was used to build systems, with proportion of frequency and time frames bands zeroed out set to 0.5 and 0.2 respectively. Additionally 6 CNN layers were added to the bottom of the network. These two changes yielded a further 1% gain. Finally more complex language models were incorporated into the system. The 4-gram LM was interpolated with a standard uni-directional RNNLM [17] and then with a 4 succeeding word RNNLM (suRNNLM) [18, 19] [4]. 4-gram lattices were rescored with this combined LM [20]. These two additional changes yielded a final WER of 18.8%. These systems were treated as the baselines for further system development.

3. System Development

Four areas were investigated to further develop the baseline ASR system presented in Table 1: the choice of lexicon and phone unit; handling of code-switching; grade dependent models; and acoustic model diversity.

3.1. Lexicon

The lexicon distributed with the challenge is based upon the CMU lexicon i.e. American English pronunciations. It was compared to two additional forms of lexicon. The first, Combilex [21] Received Pronunciation (RP) lexicon contains British English pronunciations. From Table 2, using the Baseline system in Table 1 degraded system performance compared to the CMU lexicon. This indicates that the American English pronunciations from the CMU lexicon may be a better match to the Italian children's learner speech.

Table 2: *Effect of lexicon on Dev set % WER.*

Lexicon	% WER			
	A1	A2	B1	Total
Distributed	23.4	17.4	21.8	21.2
Combilex	22.3	19.7	22.9	22.0
Graphemic	20.4	17.7	21.0	20.1

A graphemic lexicon has been shown to outperform a phonetic lexicon on non-native learner ASR, particularly for lower proficiency speakers [22]. Following [23] the graphemic set here consists of the 26 letters of the English alphabet, 2 graphemes to model all forms of hesitations, (G00, G01), plus 2 graphemes to model <unk-it>, (G02,G03), and

[4] Built with the CUED-RNNLM V1.1 toolkit [18].

<unk-de>, (G04,G05). The result in Table 2 shows that the graphemic lexicon system reduces the WER over the phonetic systems, by 1.1% absolute. Only the A2 speakers score lower with the distributed phonetic lexicon. A2 and B1 responses are similar in length but the A2 speech has a lot fewer disfluencies. This may be related to the nature of the exams, with A2 students speaking more precisely and taking greater care pronouncing words, thus being a closer match to the phonetic lexicon.

3.2. Code-switching

Although the children are being asked to speech English they fall back on their native language occasionally, for example when they can't find the right word or phrase. Table 3 shows the percentage of code-switching on a per grade level. Over 4% of words spoken by A2 and B1 speakers are marked in the reference transcriptions as code-switched words, and over 9% of A1 words. Whilst the majority of the code switching is into Italian, about 1/6th of the words are German, a reflection of the languages spoken in the Trentino region. A small proportion of all utterances (% Utt) are entirely in a code-switched language.

Table 3: *% code-switched Italian and German words and utterances only containing non-English in 2017 Train and Dev data.*

Grade	% Words			% Utt
	Italian	German	All	
A1	7.67	2.15	9.82	2.48
A2	4.01	0.86	4.87	2.84
B1	3.88	0.18	4.06	1.54
All	4.87	0.87	5.74	2.38

In the Shared Task code-switched words are eliminated from the reference at scoring so precise recognition is not needed. They do, however, need to be identified and removed from the hypothesis to avoid insertion errors. As described in Section 2 the distributed recipe models code-switched words as <unk-it>/<unk-de> which can be stripped out from the recognition hypothesis prior to scoring. The phonetic ALTA TDNN-F system, PF1, has a 18.8% WER (Table 1) and it's equivalent graphemic system, GF1, a 18.1% WER.

An alternative approach was investigated where the code-switched words were explicitly modelled in acoustic model and language model training. Each code-switched word was tagged with a language tag to identify them e.g. allora_it, ist_de. After recognition these tagged words were removed from the output hypothesis. The phonetic pronunciations for these words were taken from the German and Italian lexicons supplied with the task. To unify the phone sets all 3 lexicons were mapped to a common X-SAMPA phone set of 75 phones. A second mapping was then applied to map them to the CMU English 39 phone set. This mapping was hand derived based on the phonetic attributes of the xenophones. The Sequitur G2P [24] [5] tool was used to produce pronunciations for words missing from the lexicon. As for the baseline, a PLP GMM-HMM based HTK alignment was run to align unknown words in the transcriptions to the closest word in this extended vocabulary. Graphemic pronunciations were generated in the standard fashion. Two versions of phonetic and graphemic TDNN-F models were trained. In the first, PF2/GF2, the standard phone sets are used. For the other, PF3/GF3, language attributes were added to the phones in the code-switched words e.g.

```
ist_de   ih;G_B s;G_I t;G_E
tuo_it   t;I_B u;I_I o;I_E
```

Context dependent states within each phone are tied using decision trees [25]. The trees are able to ask questions related to the language attributes, thereby, potentially tying code-switched contexts more closely together. Different trees were observed to have been generated. A common language model was trained on the transcriptions including the code-switch tagged words. Negligible differences to the overall word error were observed for both graphemic and phonetic systems. Gains were seen in moving from PF1/GF1 to the tagged systems on A1 and B1 speech, with decreases in WER of 0.2%-0.9%. There were rises in WER from 0.5-0.7%, however, on A2 speech.

Table 4: *Recognition statistics for the 382 code-switched words in Dev.* †*Count consists of insertion of Italian or German words plus mis-recognitions of English into Italian or German.*

Sys	Recognised As		Del	Ins†	Task
	Eng	It/De			Error
GF1	108	30	244	4	27.2%
GF2	109	66	207	19	27.7%
PF1	137	1	244	0	34.6%
PF2	119	69	194	14	30.4%

The limited effect on the overall recognition performance is potentially a reflection on the scoring approach adopted. Table 4 presents statistics on the development set of how the Italian and German words were recognised. The task error measures the % of code-switched words recognised as full English words, mirroring the task scoring metric which excludes unknown, partial and foreign words. About one third of code-switched words are mis-recognised as an English word. Note, in some cases this is the actual word e.g. *ciao* occurs in the vocabulary as both an "English" word and with an Italian tag. The tagged models (GF2/PF2) are better at recognising code-switched words than the unknown models (GF1/PF1). The two graphemic models, however, yield a very similar task error as this doesn't affect their rate of mis-recognition into English. In contrast, the PF1 models tend to recognise code-switched words as English rather than unknown giving a 7% increase in task error, compared to 3% for PF2. This also shows the graphemic models for the unknown foreign words match better.

3.3. Grade dependent models

The pronunciation and grammar of learners becomes more native-like as they progress, they increase their vocabulary and reduce their code-switching. As noted in the Introduction, the exam questions asked and length and complexity of responses varies with level, from A1's few words to B1's more complex but equally disfluent speech. At test time the grade of the speaker is known so it makes sense to see if tuning the acoustic and language models to specific grades can help performance.

Two approaches to adapting the AMs to be grade dependent were investigated. In the first, the general AMs were fine-tuned using only data from a specific grade, or grades. The second approach used 1-of-K coding of the grade as an auxiliary feature for the AM. The latter approach did not yield consistent gains for any grade level, so the results presented here focus on the fine-tuning approach. Table 5 shows the impact of fine-tuning on the A2 and B1 grade data using both the 4-gram LM and the

[5] https://github.com/sequitur-g2p/sequitur-g2p.

more advanced su-RNNLM. Note no gains were obtained for the A1 grade data for either configuration. Some of the gains seen with the 4-gram decoding are lost when suRNNLM rescoring is applied.

Table 5: *Fine-tuning AMs for A2 and B1. GF2 models*

Grade Dep.	% WER			
	A2		B1	
	4g	suRNN	4g	suRNN
—	17.6	16.6	19.8	18.0
A2+B1	17.6	16.1	19.0	18.0
B1	—	—	18.5	17.4

Table 6 presents the effect on performance of interpolating the grade independent 4-gram LM with a grade dependent 4-gram LM, the interpolation weights were optimised on the development data. No improvements are seen for A2 and B1. This is probably due to the training data being biased to their longer answers, particularly with the inclusion of the written data. For A1 speakers a 0.9% drop in WER is seen. Fine-tuning and training only on the target A1 grade data were also examined for the RNNLM and suRNNLM. Neither approach, however, yielded gains over the grade independent LMs.

Table 6: *Grade independent (GI) and dependent (GD) 4-gram LMs. GF2 models.*

Grade	GI	GD	% WER
A1	1.00	—	21.5
	0.01	0.99	20.6
A2	1.00	—	17.6
	0.36	0.64	17.9
B1	1.00	—	19.8
	0.26	0.74	19.9

3.4. Acoustic model diversity

System combination of complementary systems has proven very successful in ASR [26, 27, 28]. One way to achieve this is through acoustic model diversity. Interleaved TDNN-F and long short-term memory (LSTM) configuration (TDNN-F_LSTM) [29, 30] systems were trained using lattice free MMI. Table 7 shows the performance of TDNN-F and TDNN-F_LSTM models are similar. Confusion network combination (CNC) [27] was used to combine the 2 systems which gave an improvement of about 0.6% absolute WER.

Table 7: *CNC of graphemic TDNN-F and TDNN-F_LSTM systems, with modelling of code-switched words as GF2.*

System		A1	A2	B1	Total
TDNN-F	(GF2)	19.1	16.6	18.0	18.0
TDNN-F_LSTM	(GL2)	19.6	17.4	18.0	18.3
GF2 $\oplus$ GL2		19.5	15.3	17.3	17.4

4. Evaluation Systems

During the 9 day period of the evaluation the ALTA team entered the maximum 7 systems allowed for the Closed Task. The systems all involved system combination of graphemic and phonetic systems, and TDNN-F and TDNN-F_LSTM acoustic models. During the evaluation the only feedback available to participants in terms of evaluation performance was the overall WER and numbers of substitution, insertion and deletion errors. The per grade breakdown was not available.

Table 8: *Single and combined systems on Dev and Eval*

System	Test	% WER			
		A1	A2	B1	Total
GF2	Dev	**19.1**	16.6	**18.0**	**18.0**
	Eval	20.3	14.7	17.5	17.3
PF2	Dev	20.6	17.3	18.3	18.7
	Eval	**19.5**	15.2	16.9	**16.9**
GF3	Dev	19.2	**16.5**	18.3	18.1
	Eval	19.8	**14.7**	17.3	17.0
PF3	Dev	21.1	17.1	18.5	18.9
	Eval	21.4	14.9	**16.6**	17.1
C1	Dev	19.2	15.0	16.5	16.9
	Eval	20.2	13.8	**15.5**	15.9
C2	Dev	**18.5**	**14.7**	**16.4**	**16.6**
	Eval	19.7	13.5	16.0	16.0
ROV(C1,C2)	Dev	18.8	14.9	16.4	16.7
	Eval	**19.4**	**13.4**	**15.5**	**15.7**

Table 8 shows the best single systems and combined systems on the Dev and Eval sets. As can be seen the best Eval system differs from the best Dev system at each grade and overall. The first CNC system, C1, was selected as the best grade independent system on the Eval. It was formed by combining all 4 unk-de/it systems with all 4 tagged v2 systems:

(GF1$\oplus$GL1$\oplus$PF1$\oplus$PL1)$\oplus$(GF2$\oplus$GL2$\oplus$PF2$\oplus$PL2)

The second CNC system, C2, used grade specific combination i.e. the best combination for each grade was used based on Dev set results. For A1 and B1 this was the tagged systems:

(GF2$\oplus$GL2$\oplus$PF2$\oplus$PL2)$\oplus$(GF3$\oplus$GL3$\oplus$PF3$\oplus$PL3)

and for A2 all 3 groups of systems were combined:

C1$\oplus$(GF3$\oplus$GL3$\oplus$PF3$\oplus$PL3)

Finally C1 and C2 were combined using ROVER [26]. This gave small gains on A1 and A2 and matched the C1 performance on B1, giving the lowest WER of 15.7% overall.

5. Conclusions

This paper has described the ALTA speech recognition system for the *INTERSPEECH 2020 Shared Task on Non-Native Children's Automatic Speech Recognition*. This data is especially challenging, as there is a more variety in children's speech than adult speech, limited training data, and all the standard issues of learner English. Starting from the distributed Kaldi recipe, a number of both practical refinements for handling "noisy" data, as well as algorithmic advances are presented. The final system makes extensive use of: data augmentation, in the form of speed perturbation and SpecAugment; lexical diversity using both phonetic and graphemic lexicons; acoustic model diversity; and final grade specific combination. The final combined system submitted for the Closed Task yielded a word error rate of 15.7% on the evaluation data, which was the best performing evaluation system. Additionally, the best individual system that contributed to this combined system, gave an error rate of 16.9% on the evaluation which was also lower than any other system submitted to the official evaluation.

6. References

[1] 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, no. 10, pp. 883–895, 2009.

[2] D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech and Language*, vol. 25, no. 2, pp. 282–306, 2011.

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

[4] Y. Wang, M. Gales, K. Knill, K. Kyriakopoulos, A. Malinin, R. van Dalen, and M. Rashid, "Towards Automatic Assessment of Spontaneous Spoken English," *Speech Communication*, vol. 104, pp. 47–56, 2017.

[5] K. Evanini and X. Wang, "Automated speech scoring for non-native middle school students with multiple task types," in *Proc. of INTERSPEECH*, 2013, pp. 2435—2439.

[6] 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.

[7] Y. Qian, K. Evanini, X. Wang, C. M. Lee, and M. Mulholland, "Bidirectional LSTM-RNN for Improving Automated Assessment of Non-native Children's Speech," in *Proc. of INTERSPEECH*, 2017, pp. 1417–1421.

[8] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, "Non-native children speech recognition through transfer learning," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[9] R. Gretter, M. Matassoni, K. Allgaier, S.Tchistiakova, and D. Falavigna, "Automatic assessment of spoken language proficiency of non-native children," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[10] R. Gretter, M. Matassoni, S. Bannò, and D. Falavigna, "TLT-school: a corpus of non native children speech," *CoRR*, vol. abs/2001.08051, 2020, to be published Proc. LREC, 2020. [Online]. Available: http://arxiv.org/abs/2001.08051

[11] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[12] 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 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[13] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. of INTERSPEECH*, 2018, pp. 3743–3747. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1417

[14] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. of INTERSPEECH*, 2016, pp. 2751–2755.

[15] S. 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 (for* HTK *version 3.4.1)*. University of Cambridge, 2009. [Online]. Available: http://htk.eng.cam.ac.uk

[16] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. of INTERSPEECH*. ISCA, 2019, pp. 2613–2617. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2680

[17] T. Mikolov, M. Karafiát, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. of INTERSPEECH*, 2010.

[18] X. Chen, X. Liu, Y. Qian, M. Gales, and P. Woodland, "CUED-RNNLM – An Open-Source Toolkit for Efficient Training and Evaluation of Recurrent Neural Network Lang uage Models," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[19] X. Chen, X. Liu, Y. Wang, A. Ragni, J. H. M. Wong, and M. J. F. Gales, "Exploiting future word contexts in neural network language models for speech recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 27, no. 9, pp. 1444–1454, 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2922048

[20] X. Liu, X. Chen, Y. Wang, M. J. F. Gales, and P. C. Woodland, "Two efficient lattice rescoring methods using recurrent neural network language models," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 8, pp. 1438–1449, 2016. [Online]. Available: https://doi.org/10.1109/TASLP.2016.2558826

[21] K. Richmond, R. A. J. Clark, and S. Fitt, "Robust LTS rules with the combilex speech technology lexicon," in *Proc. of INTERSPEECH*, 2009, pp. 1295–1298. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2009/i09_1295.html

[22] K. Knill, M. Gales, K. Kyriakopoulos, A. Ragni, and Y. Wang, "Use of Graphemic Lexicons for Spoken Language Assessment," in *Proc. of INTERSPEECH*, 2017, pp. 2774–2778.

[23] M. Gales, K. Knill, and A. Ragni, "Unicode-based Graphemic Systems for Limited Resources Languages," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[24] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, pp. 434–451, May 2008.

[25] S. J. Young and P. C. Woodland, "State clustering in HMM-based continuous speech recognition," *Computer Speech and Language*, vol. 8, no. 4, pp. 369–394, 1994.

[26] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1997, pp. 347–354.

[27] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*, vol. 27, 2000, pp. 78–81.

[28] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, vol. 25, no. 4, pp. 802–828, 2011. [Online]. Available: https://doi.org/10.1016/j.csl.2011.03.001

[29] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2017.

[30] Y. Wang, J. Wong, M.J.F.Gales, K. Knill, and A. Ragni, "Sequence teacher-student training of acoustic models for automatic free speaking language assessment," in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*, 2018.