Proceedings Paper:
Wang, Y., Wong, J.H.M., Gales, M.J.F. et al. (2018) Sequence teacher-student training of acoustic models for automatic free speaking language assessment. In: 2018 IEEE Spoken Language Technology Workshop (SLT). 2018 IEEE Spoken Language Technology Workshop (SLT), 18-21 Dec 2018, Athens, Greece. IEEE. ISBN: 9781538643358.

https://doi.org/10.1109/slt.2018.8639557

SEQUENCE TEACHER-STUDENT TRAINING OF ACOUSTIC MODELS FOR AUTOMATIC FREE SPEAKING LANGUAGE ASSESSMENT

Y. Wang, J. H. M. Wong, M. J. F. Gales, K. M. Knill, A. Ragni

Cambridge University Engineering Dept., Trumpington St., Cambridge CB2 1PZ, U.K.
Email: {yw396, jhmw2, mjfg, kmk, ar527}@eng.cam.ac.uk

ABSTRACT

A high performance automatic speech recognition (ASR) system is an important constituent component of an automatic language assessment system for free speaking language tests. The ASR system is required to be capable of recognising non-native spontaneous English speech and to be deployable under real-time conditions. The performance of ASR systems can often be significantly improved by leveraging upon multiple systems that are complementary, such as an ensemble. Ensemble methods, however, can be computationally expensive, often requiring multiple decoding runs, which makes them impractical for deployment. In this paper, a lattice-free implementation of sequence-level teacher-student training is used to reduce this computational cost, thereby allowing for real-time applications. This method allows a single student model to emulate the performance of an ensemble of teachers, but without the need for multiple decoding runs. Adaptations of the student model to speakers from different first languages (L1s) and grades are also explored.

***Index Terms*—** Automatic speech recognition, automatic spoken language assessment, lattice-free MMI, sequence teacher-student training, adaptation

1. INTRODUCTION

There is a high demand around the world for the learning of English as a second language. Assessment of a learner's language proficiency is a key part of learning both in measuring progress made and for formal qualifications required e.g. for entrance to university or to obtain a job. Given the high demand from English learners, it will be very difficult to train sufficient examiners and the introduction of automatic markers will be beneficial especially for practice situations. The diagram of a general automatic assessment system is shown in Figure 1 [1, 2, 3, 4]. In free speaking tests, the candidate is asked to speak for e.g. 20-60 seconds in response to a prompt. The text of the candidate's speech is unknown so an automatic speech recognition (ASR) system is used to determine what they said. Normally, most of the grader's input features are derived from the ASR transcriptions. Therefore, the ASR system performance is of great importance to the auto-marker system [5, 6].

In a multi-level testing scenario, learners can vary in proficiency from across the Common European Framework of Reference (CEFR) grades [7], from minimal proficiency (A1), through limited but effective (B1), to fully operational command of the spoken language (C2). The ASR system must handle this wide range of proficiencies for speakers across different first languages (L1s). This is a highly challenging task. First, there are large variations in accents, pronunciations, speaking rates, grammatical correctness, vocabulary and recording conditions. Second, there is limited non-native English ASR training data that is publicly available and thus it cannot be expected to cover all the variations that will be observed in deployment due to both speakers and recording conditions. Third, for training good ASR systems, high-quality transcriptions are required. The transcription quality of non-native English learner speech is often compromised, because it is difficult to understand and often contains unusual names. In [8], it was shown that the average word error rate (WER) between each pair of three professional transcription services was 23.5% for this form of speech data. Crowd-sourcing is often used for transcribing this data [9, 10]. It enables more transcriptions at the cost of lower inter-annotator agreement and more spelling errors. Another problem with the crowd-sourced transcription is that its deletion rate is normally very high, especially for poor English speakers because the transcribers tend to avoid transcribing uncertain information when they cannot understand the speech. These errors can be mitigated to a certain extent through the combination of multiple transcriptions [8].

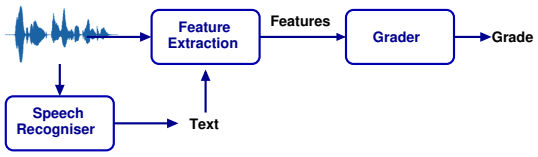

Fig. 1. Spoken language assessment auto-marker framework.

Given the challenges presented in the building of ASR systems for non-native English learner speech, it can be difficult for a single ASR system to achieve a high performance [6]. Combining an ensemble of ASR systems often gives significant performance gains over one single system, especially when the quantity of training data is limited [11]. However, if a hypothesis-level combination of the ensemble is used, such as ROVER [12], confusion network combination [13] or minimum Bayes' risk (MBR) combination [14], the computational overhead for performing recognition through the ensemble scales linearly with the number of systems and thus may be impractical for real-time applications. One possible method of reducing this computational cost is teacher-student (TS) training [15, 16, 17, 18]. This trains a single student model to emulate the combined performance of the ensemble. Only this single student model needs to be used for recognition, thereby only requiring a single decoding run. One common criterion for TS training is to minimise the Kullback-Leibler (KL) divergence between the student and teacher ensemble frame posteriors [19, 15]. Although TS

This research was partly funded under the ALTA Institute, University of Cambridge. Thanks to Cambridge Assessment English, University of Cambridge, for supporting this research and providing access to the BULATS data.

training with the frame posteriors can yield better performance than training on forced alignment hard targets, the frame-level criterion does not take into account the sequential nature of speech and sequence training can often yield significant performance gains [20]. Thus, sequence-level criteria have been introduced into the TS training framework. The aim of sequence TS training is to train a student to produce the same decoding word sequence hypothesis as the teacher ensemble. In [16], the KL divergence between the word sequence posteriors from the student and a sequence-trained teacher ensemble was used as the criterion. The teachers in the ensemble were first trained by a cross-entropy (CE) criterion and then further trained by a maximum mutual information (MMI) criterion or a state-level MBR (sMBR) criterion [21, 20]. The sequence TS training in [16] was extended in [18] to the lattice-free MMI (LF-MMI) framework, proposed in [22].

In this paper, sequence-level TS trained systems are explored for non-native English learner data. A lightly-supervised approach is used to mitigate the impact of the errors in the crowd-sourced transcriptions on the training of the LF-MMI acoustic model. Sequence-level TS training is implemented within a LF framework, allowing the student to achieve a similar performance to that of an ensemble with only a single decoding run. Adaptation of the student model to different L1s and grades is also explored.

The rest of the paper is organised as follows. Section 2 describes the LF-MMI acoustic model. Section 3 introduces the sequence-level TS training implementation, which can be used within a LF framework. Section 4 gives the experimental setup and results. Section 5 draws the conclusions.

2. LATTICE-FREE MMI ACOUSTIC MODEL

The aim of MMI training is to maximise the posterior probability of the correct utterance while minimising the probability of all other transcriptions, by minimising the objective function:

$$\begin{aligned}\mathcal{F}_{\text{MMI}} &= -\sum_{r} \log P\left(\mathbf{s}_r^*|\mathbf{O}_r, \boldsymbol{\Theta}\right) \\ &= -\sum_{r} \sum_{\mathbf{s}_r \in \mathcal{G}} \delta\left(\mathbf{s}_r, \mathbf{s}_r^*\right) \log P\left(\mathbf{s}_r|\mathbf{O}_r, \boldsymbol{\Theta}\right) \quad (1) \\ &= -\sum_{r} \log \frac{P\left(\mathbf{s}_r^*\right) p^{\kappa}\left(\mathbf{O}_r|\mathbf{s}_r^*, \boldsymbol{\Theta}\right)}{\sum\limits_{\mathbf{s}_r' \in \mathcal{G}} P\left(\mathbf{s}_r'\right) p^{\kappa}\left(\mathbf{O}_r|\mathbf{s}_r', \boldsymbol{\Theta}\right)} \quad (2)\end{aligned}$$

where $\delta\left(\cdot\right)$ represents the Kronecker delta function. $\mathbf{s}_r^*$ is correct state sequence of utterance r, $\boldsymbol{\Theta}$ represents model parameters. $\mathcal{G}$ represents the set of all state paths in the lattice, or phone graph in the LF implementation. Here, $p^{\kappa}\left(\mathbf{O}_r|\mathbf{s}_r, \boldsymbol{\Theta}\right)$ is the acoustic score with an acoustic scaling factor, κ. Ideally, the sum in the denominator should be taken over all possible sequences. This can be computationally expensive, however, when the hypothesis space is large. To reduce this computational cost, it is common in practice to limit the sum to consider only the most likely hypotheses, determined by pruning a lattice generated by a decoding pass with the training data, using an initial model. In conventional DNN based hybrid model training, the neural network is first trained using a CE criterion to provide the initial model . The CE-trained neural network is then used to generate the hypothesis lattice for each utterance, which is then used to compute the gradient for the MMI criterion of (2).

The disadvantages of this conventional MMI training approach are twofold. First, the MMI training requires a CE-trained initial model. Second, it is very expensive to generate and store the denominator lattice for each training utterance. The LF-MMI implementation in [22] avoids the need for CE initialisation and lattice generation. It calculates the denominator by directly applying forward-backward computations [23, 24] on an unpruned denominator graph on GPU hardware. For speeding up the computations, a phone-level language model, instead of a word-level language model, is used to generate a phone graph. The phone-level language model significantly reduces the number of state sequences that needs to be represented in the lattice, thereby reducing the computational cost required to perform forward-backward computations. A pruned 4-gram phone language model is normally trained using a phonetic decomposition of the training data transcriptions

To further reduce the computational cost, a 30ms frame rate rather than a 10ms frame rate is often used. When using a 10ms frame rate, the conventional 3-state HMM, shown in Figure 2, requires a minimum of 30ms to traverse. When using a 30ms frame rate, a simpler 2-state HMM topology, shown in Figure 2, can instead be used, to preserve the minimum of 30ms required to traverse a phone. In this 2-state topology, the first emitting state does not have a self-loop and can be activated only once for each occurrence of the sub-word units. Only the second emitting state has a self-loop connection. This topology is similar to the one used in the Connectionist Temporal Classification (CTC) approach [25]. This simpler topology results in fewer states and a smaller output layer, thus further reducing computation. LF-MMI systems have been shown to give better or comparable performance when compared with their lattice-based counterparts over a range of tasks [22, 26, 27, 28]. However, the LF-MMI systems are very sensitive to errors in the transcriptions [22]. This problem is further exacerbated when there is a high deletion rate in the training transcriptions, as is often the case with crowd-sourced transcriptions. High deletion rate, however, does not interact well with powerful, sharp, acoustic model as what is normally used for building LF-MMI systems [22, 26].

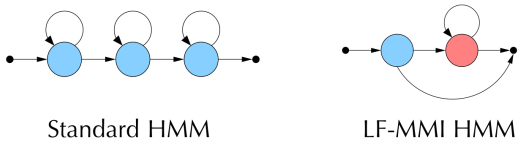

Fig. 2. HMM topology used in standard neural network models and LF-MMI models.

3. SEQUENCE TEACHER-STUDENT TRAINING

Rather than just using a single model, significant performance gains can often be obtained by combining together an ensemble of multiple models [12]. If the models have a diversity of different behaviours, then errors may occur at different locations between the hypotheses of the models. This may allow the models to correct for each other's errors. In this paper, diversity is introduced into the ensemble by beginning training from different random initialisations [16]. Ensemble methods may be especially beneficial when the quantity of training data is limited, such as the present situation of using non-native English learner speech. The ensemble of models can be combined at the hypothesis level, using methods such as ROVER [12], confusion network combination [29] and MBR combination [14]. However, these can be computationally expensive when performing recognition, as a separate decoding run is required for each of the models. This may hinder the application of ensemble methods to real-time automatic assessment systems.

TS training is one possible method that can be used to alleviate the computational cost of using an ensemble for recognition. In TS training, a single student model is trained to emulate the combined performance of the ensemble. Only this single student model then needs to be used for recognition and assessment, thereby reducing the computational cost. One possible method of training the student is to minimise the KL-divergence between the frame-level state posteriors of the student and teachers [19]. This criterion only propagates frame-level posterior information from the teacher ensemble to the student, and may not adequately capture the sequential nature of speech data. Instead, a sequence-level criterion can be used, by minimising the KL-divergence between the state sequence posteriors of the student and teachers [16],

$$\mathcal{F}_{\text{TS}} = -\sum_{r}\sum_{\mathbf{s}_r\in\mathcal{G}} P\left(\mathbf{s}_r\middle|\mathbf{O}_r,\widehat{\boldsymbol{\Phi}}\right)\log P\left(\mathbf{s}_r|\mathbf{O}_r,\boldsymbol{\Theta}\right), \quad (3)$$

where $\widehat{\boldsymbol{\Phi}}$ represents the teacher model parameters and $\boldsymbol{\Theta}$ represents the student model parameters. In this work, an ensemble is used as the teacher. There are several possible methods of combining the teachers in the ensemble to obtain the sequence posterior targets used to train the student. A simple method is to take a sum combination of the sequence posteriors [16]:

$$P\left(\mathbf{s}_r\middle|\mathbf{O}_r,\widehat{\boldsymbol{\Phi}}\right) = \sum_{m=1}^{M}\lambda_m P\left(\mathbf{s}_r|\mathbf{O}_r,\boldsymbol{\Phi}^m\right). \quad (4)$$

where M is the number of teachers in the ensemble. Here, λ_m and $\boldsymbol{\Phi}^m$ represent the interpolation weights and model parameters of teacher m, respectively. This form of targets can be interpreted as a Monte Carlo approximation to Bayesian inference of the state cluster sequence.

The gradient of the criterion in (3) is calculated as

$$\frac{\partial\mathcal{F}_{\text{TS}}}{\partial\log P\left(s_t|\mathbf{o}_t,\boldsymbol{\Theta}\right)} = \kappa\left[P\left(s_t|\mathbf{O}_r,\boldsymbol{\Theta}\right) - P\left(s_t\middle|\mathbf{O}_r,\widehat{\boldsymbol{\Phi}}\right)\right]. \quad (5)$$

This gradient can be efficiently computed using a forward-backward pass over each denominator phone graph produced by the student and teacher models. In this paper, a LF implementation of the gradient computation is used as that in [18], allowing for a simple integration with existing LF-MMI training implementations and models.

4. EXPERIMENTS

4.1. Training and evaluation data sets

Experiments were conducted using the data from the Business Language Testing Service (BULATS) Online Speaking Test of Cambridge English Language Assessment [30]. The BULATS test comprises 5 sections: A. responses to short questions; B. read aloud sentences; C-E. free speaking responses with a maximum length of 60 (sections C and D) or 20 seconds (5 parts of section E). The ASR training set, *TRN*, consists of 334 hours of BULATS data, which mainly covers 28 L1s[1] and the 5 CEFR grades ranging from A1 to C2. C1 and C2 grades are merged due to a lack of C2 speakers. The distributions of the L1s and grades are shown in Figure 3, respectively. Transcriptions that combined two crowd-sourcing transcriptions using the approach in [8] were available for this data set.

Two evaluation data sets were used. The first one, *EVL.trans*, was used for evaluating the quality of the transcriptions. This data set was also used in [8] for the same purpose. It contains 88 speakers in 10 hours of speech. All are Indian Gujarati native speakers, and are roughly evenly distributed across CEFR grade range [7]. Each utterance has been transcribed by crowd-sourcers. Additionally, each utterance has also been transcribed by professional services [8]. This allows us to evaluate the quality of transcriptions on the professional transcriptions using this data set.

The second evaluation data set, *EVL.asr*, was used for assessing the performance of ASR systems. It consists of 226 speakers from Polish, Arabic, Vietnamese (Viet.), French, Thai and Dutch L1s. The speakers are roughly evenly distributed across the CEFR grade range [7]. To assess the performance of the ASR systems, word error rates (WERs) were calculated for Sections C, D and E, containing a total of 14 hours of spontaneous speech. This data set has been transcribed by native English speaking transcribers [31, 6].

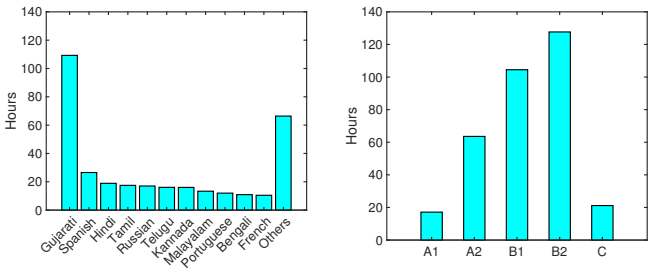

Fig. 3. L1 (left) and grade (right) distributions of TRN.

4.2. Lightly-supervised training

As discussed in Section 2, LF-MMI systems are highly sensitive to transcription errors, especially deletions. Given this, the solution proposed in this paper to reduce the impact of the poor quality of the crowd-sourced transcriptions is to use a lightly-supervised (LS) approach. LS approaches are often used to produce better transcriptions than the original transcriptions [32, 33, 34]. The idea is to use an acoustic model and a biased LM to recognise the training data and generate the hypotheses which are then selected and used as the training transcriptions. In order to produce high quality transcriptions using a strong system, in this paper, a joint stacked hybrid DNN and LSTM system is used for the LS transcription generation.

The diagram of the joint system is shown in Figure 4. This system was shown to give good performance for a range of downstream tasks in the BULATS test in [6]. A bottleneck DNN (BN-DNN) was first trained. The BN-DNN had a $720\times1000^4\times39\times1000\times6000$ structure, where the BN feature size is 39. The input feature vector of the BN-DNN consisted of 9 consecutive frames of 40-dimensional log Mel-filter bank features plus their delta. Thus the dimension of the feature vector was 720. The bottleneck features were appended with a 52-dimensional PLP$+\Delta+\Delta^2+\Delta^3$ feature vector. An HLDA transform was applied to the PLP features and a global semi-tied transform was applied to the BN features, reducing the dimension of the combined feature vector from 91 to 78. This feature vector was used to train both DNN and LSTM Hybrid acoustic models with a context window of [-4,4]. The hybrid systems were implemented using the Kaldi toolkit [35], the DNN had a $702\times1000^5\times8949$ structure with 8949 tri-graphemic state outputs. A graphemic lexicon was used because it was shown to give a better performance than a phonetic lexicon for this task [36]. The LSTM had 2 hidden layers, each with 1000 memory cells and 500 recurrent projection units [37]. Cross entropy a state-level MBR (sMBR)-based sequence

[1]There are 75 L1s in total but most have only 1 or 2 speakers.

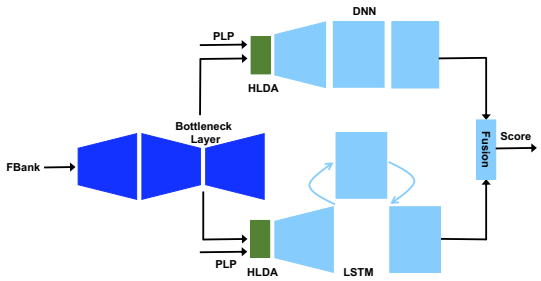

Fig. 4. Joint decoding system that is used for generating the lightly-supervised (LS) transcriptions.

training were applied [21]. The models were trained on the combined crowd-sourced transcriptions. An in-domain LM was trained on 1.83M words from the combined crowd-sourced transcriptions of the training data, using the SRILM toolkit [38]. This LM was then interpolated with a general LM that was trained on Broadcast News English (BNE) [39] using an interpolation weight of 0.83 for the BULATS LM and a weight of 0.17 for the BNE LM. The choice of a large interpolation weight for the TRN data allows the LM to be used as a biased LM. Thus, the data in TRN can be decoded using this system to produce LS transcriptions for training the LF-MMI system.

Trans.	%WER						RTF
	A1	A2	B1	B2	C	Overall	
CWD	40.3	31.3	27.5	18.6	15.9	25.1	–
Joint	39.7	31.4	27.4	19.1	17.8	25.6	3.1

Table 1. %WER on EVL.trans for different grades for crowd-sourced (CWD) transcriptions and transcriptions produced by the joint DNN and LSTM hybrid system. The decoding real-time factor (RTF) is also calculated for the joint system.

Table 1 shows the WERs on EVL.trans for different grades for the crowd-sourced transcriptions and the transcriptions produced by the joint system using as the reference the professional transcriptions. It can be seen that the joint DNN and LSTM system yields a comparable overall WER to the crowd-sourcers. As the English proficiency of the speakers improves, the performance gap increases. For A1 speakers, whose English is the most difficult to transcribe, the transcriptions produced by the joint DNN and LSTM system is better than the crowd-sourced transcriptions. Also, for A1 speakers, the crowd-sourced transcriptions have a deletion error rate of 14.3% while the deletion error rate from the joint system is only 12.3%. Thus, the transcriptions from the joint system are of a higher quality and have a lower deletion rate than the crowd-sourced transcriptions for poorer speakers. Moreover, it is expected that the transcriptions from the ASR systems are more consistent than the crowd-sourced transcriptions across the utterances where there are large variations in the accents and audio quality. The table also shows that the decoding real-time factor (RTF) required by the joint system is 3.1[2], which makes it challenging to deploy in a real-time system.

Two 9-layer interleaved time-delay DNN (TDNN) and long-short term memory (LSTM) LF-MMI systems were built using the Kaldi toolkit [35] using crowd-sourced transcriptions and the LS transcriptions from the joint system, respectively. The TDNN-LSTM models had the same structure as that in [28]. They had 6 TDNN [40] layers with 600-dimensional ReLU units and 3 LSTM layers with 512 cells with 128 recurrent and non-recurrent projections [37]. The acoustic model features were 40-dimensional log Mel-filter bank features normalised using speaker level mean and variance normalisation. Around 2500 context-dependent graphemic states were used as output targets, which is much fewer than that of the joint DNN and LSTM system. The LM is the same as that used in the joint system.

Table 2 shows the WER on EVL.asr for different grades using the two LF-MMI systems. It can be seen that the LF-MMI trained on the LS transcriptions performs consistently better than that trained on the crowd-sourced transcriptions across different grades. Overall, it gives about 2% absolute WER improvement, with the performance gap increasing as the grade becomes worse. For A1 speakers, the LF-MMI system trained on the LS transcriptions gives about 5% absolute WER reduction over that trained on the crowd-sourced transcriptions. This implies that the LF-MMI system trained on the LS transcriptions is significantly more robust for data that is highly challenging to transcribe. Compared with the joint system, the decoding RTF required by the LS trained LF-MMI system is only 0.6, which is about 5 times faster than that of the joint system.

System	%WER						RTF
	A1	A2	B1	B2	C	Overall	
CWD	49.2	39.9	29.7	27.8	23.8	30.4	0.6
LS	44.4	36.9	27.4	26.3	22.4	28.3	

Table 2. %WER on EVL.asr for different grades for the LF-MMI TDNN-LSTM systems trained on crowd-sourced (CWD) and lightly-supervised (LS) transcriptions.

4.3. Teacher-student training

As shown above, the LF-MMI system is much less computationally expensive to use for recognition than the joint DNN and LSTM system. It does, however, have a poorer WER than the joint DNN and LSTM system shown in Figure 4, which gives a WER of 25.9% on EVL.asr. This is because the joint system can leverage the combination of a DNN model and a LSTM model. In this paper, ensemble method is used to improve the performance of the LF-MMI system. Ensemble methods previously have been shown to give significant gains over single systems [12, 41, 16]. Ensemble methods can also be applied to the LF-MMI system to potentially provide performance gains. In this experiment, an ensemble was generated by training 3 TDNN-LSTM LF-MMI systems, beginning from different model parameter random initialisations [16]. These are trained toward the 1-best LS transcriptions provided by the joint DNN and LSTM system. Hypothesis-level combination was performed using minimum Bayes' risk (MBR) combination. However, hypothesis-level combination can be computationally expensive, as a separate decoding run is required for each member of the ensemble. This can present a hindrance when aiming to develop a real-time automatic assessment system. TS training can be used to reduce this computational cost. A LF implementation of sequence-level TS training was used to train a single student model to emulate the combined performance of the ensemble. The combined state sequence posteriors of the ensemble, which was given in (4), was used as the training target. The student used the same TDNN-LSTM structure as each of the teachers in the ensemble, and therefore should have a decoding RTF that is similar

[2]The real-time factors were calculated based on threads rather than cores.

to each of the teachers in the ensemble. Equal interpolation weights were used for both MBR combination and TS training.

Table 3 compares the ensemble performances using these different combination methods. The results suggest that the ensemble can outperform a single LF-MMI system. Using hypothesis-level MBR combination, the ensemble consistently performs better than the single LF-MMI system across the grades and on average it gives 1.6% absolute performance gain. For each grade, the performance gain from the ensemble combination is similar. However, the RTF of the MBR ensemble combination is about 3 times of that of the single system, mainly due to the 2 decoding runs that are required for the additional members of the ensemble. The last row of Table 3 shows the performance of the student model. It can be seen from the table that, the student yields a slightly better performance than the MBR combination of the ensemble. By looking at the performance for each grade, the largest WER improvement, which has a 1% absolute WER reduction, is from the A1 speakers. This suggests that the student model, trained toward the state sequence posterior targets of the ensemble, may be more robust to the erroneous transcriptions than the LF-MMI models that are trained toward 1-best targets. When using TS training to combine an ensemble, only a single student model is used for recognition. As such, the single student model has a decoding RTF of 0.6, similar to that of a single LF-MMI system. Furthermore, the student model has a decoding RTF that is about 5 times faster than the joint DNN and LSTM system[3], with only a slight degradation in the WER performance. Thus, this student model may be more suitable for real-time deployment.

System	%WER						RTF
	A1	A2	B1	B2	C	Overall	
Single	44.4	36.9	27.4	26.3	22.4	28.3	0.6
Ensemble+MBR	42.9	35.4	25.2	24.6	21.2	26.7	1.8
Ensemble+TS	41.9	35.1	25.2	24.2	21.0	26.4	0.6

Table 3. %WER on EVL.asr for different grades and real-time factors (RTF) for single LF-MMI system, MBR combination of LF-MMI ensemble and TS trained LF-MMI system with a teacher ensemble.

The joint DNN and LSTM system has a slightly better performance than the combined LF-MMI ensemble. As such, it may be beneficial to instead use this joint system as the teacher for teacher-student training. This may allow the good performance of the joint DNN and LSTM system to be leveraged upon directly, while maintaining the fast decoding RTF of the student model with a TDNN-LSTM structure. The differences between the sets of state clusters, and frame rates of the joint system and TDNN-LSTM model make it difficult, however, for the sequence-level teacher-student training method that has been described in Section 3 to be used directly. It may be possible to generalise the teacher-student training framework, using methods such as that described in [42], to allow for these differences between models. This may be an interesting direction for future research.

4.4. L1 and grade adaptation

The systems considered thus far have been trained on data from all L1s and grades. These systems are therefore fairly general. However, speech from speakers with different L1s or estimated grades may have significantly different characteristics. It may be possible to adapt a general system toward the L1 or grade of the speaker who is being assessed. This is in some ways similar to speaker adaptation methods [43].

It is assumed in this paper that the L1 specified by each speaker is correct. In practice, the L1 could be predicted using methods such as in [44]. The grade which a speaker is estimated to fall into can be obtained by using a first pass automatic assessment using the general system. In the initial work in this paper, however, it is assumed that the grade for each speaker is known before-hand, and is taken as that which has been assigned by professional human examiners and is provided with the EVL.asr data. It is also assumed that the L1 and grade ranges for the test speakers are the same as those for the speakers in the training set. Data used for adaptation was obtained from within TRN, and its distribution over L1s and grades is shown in Table 4. The total amount of adaptation data (about 21 hours) is about 6% of the amount of TRN (about 334 hours). The columns of the table give the amount of data for each L1 over all grades. Except for French, which has about 10 hours of data, there is very limited data for other L1s, especially for Thai which only has about 0.92 hours of data. The amount of data for each grade varies for each L1. In general, B1 and B2 take up most of the data for each L1, and A1 has the least amount of data with only 0.77 hours in total. Unlike the TRN data, the data in the EVL.asr set has been designed such that there is a fairly even distribution of the quantity of data across all L1s and grades.

Hrs	Polish	Arabic	Viet.	French	Thai	Dutch	Total
A1	0.18	0.09	0.10	0.33	0.07	0.00	0.77
A2	0.45	0.21	0.13	1.14	0.29	0.00	2.22
B1	1.20	0.52	1.24	4.16	0.38	0.16	7.66
B2	0.75	0.61	1.54	4.29	0.13	1.29	8.61
C	0.13	0.17	0.18	0.52	0.05	0.98	2.03
Total	2.71	1.60	3.19	10.44	0.92	2.43	21.29

Table 4. %L1 and grade band distribution of the adaptation data from TRN. The L1s in EVL.asr are considered.

In this experiment, the student model, trained toward the LF-MMI ensemble on all of the TRN data, is used as the initial general acoustic model. Two methods are considered to adapt this general model to each L1 or grade. The first uses LF-MMI training with only the adaptation data for each L1 or grade, where the targets are obtained as the 1-best LS transcriptions. This form of targets is referred to as "LS-1best". The second uses sequence-level TS training, again on only the adaptation data for each L1 or grade, where the targets are obtained as the state sequence posteriors from the LF-MMI ensemble. Unlike the 1-best LS transcriptions, this form of targets may capture the uncertainty about what the correct transcription of the adaptation data utterances should be, represented within a lattice or graph. This form of targets is therefore referred to as "En-graph". The propagation of uncertainty about the transcriptions may be particularly beneficial for utterances that are more difficult to transcribe and therefore are more prone to errors in the 1-best transcriptions. These two methods can be viewed as realisations of the same KL-divergence criterion, with different forms of targets, as shown in (1) and (3). Using either form of targets, the general student acoustic model is adapted by performing one epoch of training. The learning rates used for the different L1s or different grades are the same.

Table 5 first considers adaptation to the L1 of the speaker. The adapted systems for each L1 are tested on the speakers belonging to that L1 from the EVL.asr test set. The results suggest that L1

[3]The decoding configurations for both the joint and LF-MMI systems were not optimised and the standard configurations in Kaldi were used for both systems.

adaptation using both forms of targets yields consistent gains across all L1s. Using the LS-1best targets, the adaptation yields good performance gains for Polish and Vietnamese, and the WERs for these two L1s are reduced by about 5% relative. Adapting to Thai, which has the least amount of adaptation data, yields the least performance gain, which is only a 1% relative WER reduction. Overall, L1 adaptation using LS-1best targets drops the WER from 26.4% to 25.6%. This performance is similar to that of the joint DNN and LSTM system, but has a much smaller decoding RTF, assuming that the L1 of the test speaker is known. L1 adaptation using the En-graph targets does not yield as much performance gains as when using the LS-1best targets. This may be because the initial general student model has already been trained toward the same targets from the ensemble, so there may be little more that can be gained.

System	Target	%WER						
		Polish	Arabic	Viet.	French	Thai	Dutch	Overall
TS	En-graph	20.8	31.4	32.2	22.4	30.0	20.8	26.4
+Adapt	LS-1best	19.8	30.5	30.6	22.0	29.7	20.2	25.6
	En-graph	20.3	30.8	30.9	21.9	29.9	20.4	25.8

Table 5. %WER on EVL.asr for L1-level adaptation of the TS trained system.

Table 6 next considers adapting the acoustic model to the estimated grade that the speaker is assessed as having. In this initial work, it is assumed the estimated grade is taken as that which has been assigned by professional human examiners, provided with the TRN and EVL.asr data. Starting again from the general student model, adaptation data for an estimated grade from all L1s are used to adapt the acoustic model to that grade. The adapted systems for each grade are tested on the speakers with the same estimated grade from the EVL.asr test set. Consistent gains are seen to be obtained using both forms for targets for grade adaptation. The gains obtained from grade adaptation, however, are less than those from L1 adaptation in Table 5. This may indicate that the variability of speech between different L1s is greater than the variability within a L1 between different estimated grades. For grade adaptation using the LS-1best targets, the performance gains appear to increase as the speakers' English proficiency improves from A1 to C. This may be related to the observation in Table 1 that the quality of the LS transcriptions improves with the grade of the speaker. On the other hand, the gain for A1 speakers after adaptation is significantly greater when using the En-graph targets, than when using the LS-1best targets. This suggests that the information about the transcription uncertainty that is captured within the En-graph targets may be especially beneficial when the quality of the 1-best targets is poor.

System	Target	%WER					
		A1	A2	B1	B2	C	Overall
TS	En-graph	41.9	35.1	25.2	24.2	21.0	26.4
+Adapt	LS-1best	41.8	34.9	24.8	23.8	20.5	26.0
	En-graph	41.3	35.0	25.0	23.9	20.7	26.1

Table 6. %WER on EVL.asr for grade-level adaptation of the TS trained system.

5. CONCLUSIONS

This paper explores the building of a high performance ASR system for non-native spontaneous English learner data that is suitable for real-time deployment. By leveraging upon a sequence-level teacher-student training approach, a LF-MMI system can emulate the performance of the combination of an ensemble while only requiring a single decoding run, and with a RTF of about 0.6. This system can also be adapted to speakers from different L1s and grades using both the 1-best word sequence and the state sequence posteriors as the targets. The adaptation can yield performance gains of about 5% relative in the best-case scenario.

6. REFERENCES

[1] K. Zechner et al. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895, 2009.

[2] S. M. Witt and S. J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2):95–108, 2000.

[3] A. Metallinou and J. Cheng. Using deep neural networks to improve proficiency assessment for children English language learners. In *Proc. Interspeech*, pages 1468–1472, 2014.

[4] R. C. van Dalen, K. M. Knill, and M. J. F. Gales. Automatically grading learners' English using a Gaussian process. In *Proc. SLaTE*, pages 7–12, 2015.

[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 *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pages 294–299. IEEE, 2014.

[6] K. M. Knill et al. Impact of ASR performance on free speaking language assessment. to appear in Proc. Interspeech, 2018.

[7] Council of Europe. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, 2001.

[8] R. C. van Dalen et al. Improving multiple-crowd-sourced transcriptions using a speech recogniser. In *Proc. ICASSP*, pages 4709–4713, 2015.

[9] R. Snow et al. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. EMNLP*, pages 254–263, 2008.

[10] G. Parent and M. Eskenazi. Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. In *Proc. Interspeech*, pages 3037–3040, 2011.

[11] T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems (Proceedings of MCS 2000)*, pages 1–15. Springer-Verlag, LNCS, 2000.

[12] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. ASRU Workshop*, pages 347–354, 1997.

[13] G. Evermann and P. C. Woodland. Posterior probability decoding, confidence estimation and system combination. In *Proc. Speech Transcription Workshop*, volume 27, pages 78–81, 2000.

[14] H. Xu et al. Minimum Bayes risk decoding and system combination based on a recursion for edit distance. *Computer Speech & Language*, 25(4):802–828, 2011.

[15] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *Proc. NIPS*, Montréal, Canada, 2014.

[16] J. H. M. Wong and M. J. F. Gales. Sequence student-teacher training of deep neural networks. In *Proc. Interspeech*, pages 2761–2765, 2016.

[17] J. H. M. Wong and M. J. F. Gales. Multi-task ensembles with teacher-student training. In *Proc. ASRU Workshop*, pages 84–90, 2011.

[18] N. Kanda, Y. Fujita, and K. Nagamatsu. Investigation of lattice-free maximum mutual information-based acoustic models with sequence-level Kullback-Leibler divergence. In *Proc. ASRU Workshop*, pages 69–76, 2017.

[19] J. Li et al. Learning small-size DNN with output-distribution-based criteria. In *Proc. Interspeech*, pages 1910–1914, 2014.

[20] K. Veselỳ et al. Sequence discriminative training of deep neural networks. In *Proc. Interspeech*, pages 2345–2349, Aug 2013.

[21] M. Gibson and T. Hain. Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition. In *Proc. Interspeech*, pages 2406–2409, 2006.

[22] D. Povey et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Proc. Interspeech*, pages 2751–2755, 2016.

[23] V. Valtchev et al. Lattice-based discriminative training for large vocabulary speech recognition. In *Proc. ICASSP*, volume 2, pages 605–608, 1996.

[24] P. C. Woodland and D. Povey. Large scale discriminative training of hidden markov models for speech recognition. *Computer Speech & Language*, 16(1):25–47, 2002.

[25] A. Graves et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*, pages 369–376, 2006.

[26] V. Peddinti et al. Low latency acoustic modeling using temporal convolution and LSTMs. *IEEE Signal Processing Letters*, 25(3):373–377, 2017.

[27] T. Alumäe et al. The 2016 BBN georgian telephone speech keyword spotting system. In *Proc. ICASSP*, pages 5755–5759, 2017.

[28] Y. Wang et al. Phonetic and graphemic systems for multi-genre broadcast transcription. In *Proc. ICASSP*, pages 5899–5903, 2018.

[29] L. Mangu, E. Brill, and A. Stolcke. Finding consensus among words: lattice-based word error minimization. In *Proc. Eurospeech*, 1999.

[30] BULATS. Business Language Testing Service. Available: `http://www.bulats.org/computer-based-tests/online-tests`.

[31] A. Caines, D. Nicholls, and P. Buttery. Annotating errors and disfluencies in transcriptions of speech. Technical Report UCAM-CL-TR-915, University of Cambridge, Computer Laboratory, Dec 2017.

[32] L. Lamel, J. Gauvain, and G. Adda. Lightly supervised and unsupervised acoustic model training. *Computer Speech & Language*, 16(1):115–129, 2002.

[33] Y. Long et al. Improving lightly supervised training for broadcast transcription. In *Proc. Interspeech*, pages 2187–2191, 2013.

[34] P. Lanchantin et al. Selection of multi-genre broadcast data for the training of automatic speech recognition systems. In *Proc. Interspeech*, pages 3057–3061, 2016.

[35] D. Povey et al. The Kaldi speech recognition toolkit. In *Proc. ASRU Workshop*, 2011.

[36] K. M. Knill et al. Use of graphemic lexicons for spoken language assessment. In *Proc. Interspeech*, pages 2774–2778, 2017.

[37] H. Sak, A. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proc. Interspeech*, pages 338–342, 2014.

[38] A. Stolcke. SRILM-an extensible language modeling toolkit. In *Proc. ICSLP*, pages 901–904, 2002.

[39] J. Garofolo, J. G. Fiscus, and W. M. Fisher. Design and preparation of the 1996 Hub-4 broadcast news benchmark test corpora. In *Proc. DARPA Speech Recognition Workshop*, pages 15–21, 1997.

[40] V. Peddinti, D. Povey, and S. Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proc. Interspeech*, pages 3214–3218, 2015.

[41] L. Deng and J. C. Platt. Ensemble deep learning for speech recognition. In *Proc. Interspeech*, pages 1915–1919, 2014.

[42] J. H. M. Wong and M. J. F. Gales. Student-teacher training with diverse decision tree ensembles. In *Proc. Interspeech*, pages 117–121, 2017.

[43] M. J. F. Gales. Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition. *Computer Speech and Language*, 12(2):75–98, 1998.

[44] K. Kyriakopoulos, M. J. F. Gales, and K. M. Knill. Automatic characterisation of the pronunciation of non-native English speakers using phone distance features. In *Proc. SLaTE*, pages 59–64, 2017.