LEARNING BETWEEN DIFFERENT TEACHER AND STUDENT MODELS IN ASR

Jeremy H. M. Wong

Microsoft Corporation
One Microsoft Way, Redmond, WA 98052

Mark J. F. Gales and Yu Wang

Department of Engineering, University of Cambridge
Trumpington Street, CB2 1PZ Cambridge, England

ABSTRACT

Teacher-student learning can be applied in automatic speech recognition for model compression and domain adaptation. This trains a student model to emulate the behaviour of a teacher model, and only the student is used to perform recognition. Depending on the application, the teacher and student may differ in their model types, complexities, input contexts, and input features. In previous works, it is often shown that learning from a strong teacher allows the student to perform better than an equivalent model trained with only the reference transcriptions. However, there has not been much investigation into whether a particular form of teacher is appropriate for the student to learn from. This paper aims to study how effectively the student is able to learn from the teacher, when differences exist between their designs. The AMI meeting transcription and MGB-3 television broadcast audio tasks are used in this analysis. Experimental results suggest that a student can effectively learn from a more complex teacher, but may struggle when it lacks input information. It is therefore important to carefully consider the design of the student for each application.

Index Terms— Teacher-student, acoustic model, feature vector, automatic speech recognition, Gaussian mixture model

1. INTRODUCTION

Teacher-student learning [1] has been applied in Automatic Speech Recognition (ASR), for both model compression [2] and domain adaptation [3]. In model compression, a compact student model is trained to emulate a larger teacher [2] or ensemble of multiple teachers [4]. Only this single student needs to be used to perform recognition, thereby reducing the computational cost. In domain adaptation, a student that uses input features from one domain is trained to emulate a teacher that uses input features from another time aligned domain. The features used for the student are often easier to obtain when performing recognition than those used by the teachers, but may result in degraded performance if used with standard training methods. Teacher-student learning can allow a student that uses these different features to behave similarly to the teacher.

It is often the case that the teacher and student are designed to be different. Work in [2] trains a feed-forward Deep Neural Network (DNN) student with narrow hidden layers to emulate a DNN teacher with wider hidden layers. Work in [5] trains a DNN student to emulate a Recurrent Neural Network (RNN) teacher. Work in [3] trains a student that takes far-field input features to emulate a teacher that uses input features from a close-talking microphone. In these studies, the performance of the student is often compared to the performance of an equivalent model that is trained using standard cross-entropy or sequence training methods. However, a comparison is seldom made with a student that is trained toward a teacher with the same topology and features, to assess the impact of the differences between the teacher and student.

This paper aims to study how the differences between the teacher and student models affect the ability of the student to learn from the teacher. It is interesting to question how different the student can be made from the teacher, while still being able to learn effectively. Work in [6] performs teacher-student learning between models with different decision trees, and shows that the student requires a sufficiently large decision tree to effectively learn from an ensemble of teachers. This suggests that careful consideration may be prudent when designing the student. This paper assesses the ability to effectively propagate information between the teacher and student, when the models and features differ.

2. MODEL DIFFERENCES

There are many ways in which the teacher and student can differ, such as by using different decision trees [6]. This paper considers four separate differences that can exist between the teacher and student.

2.1. Model complexity

Teacher-student learning can be used to compress a large teacher into a smaller student [2], to reduce the computational cost of performing recognition. This smaller student is designed to have fewer parameters than the teacher. This can be achieved by, for example, having fewer or narrower hidden layers. The resulting student requires less memory to store and is faster to use for recognition. However, the trade-off is that reducing the number of parameters may diminish the model's ability to capture complex behaviours. The greater flexibility of a larger model often yields an improved performance when sufficient training data is available. It is therefore interesting to access how well a smaller model can learn to emulate the more complex behaviour of a larger model.

2.2. Model type

Many different model types can be used for ASR. These include Hidden Markov Model (HMM) [7], encoder-decoder [8], RNN transducer [9], and connectionist temporal classification [10] models. The study in this paper is restricted to HMM-based models, and extending teacher-student learning to propagate information across different model types may be an interesting future research direction. Even when using HMMs, several different types of acoustic models can be used to compute the observation likelihoods. Two

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

such examples are the Gaussian Mixture Model (GMM) [11] and Neural Network (NN) [7].

GMM acoustic models used to be the state-of-the-art, until recently when NNs came into fashion. Using a NN as the acoustic model for an HMM is commonly referred to as a hybrid model. NNs have several advantages over GMMs. NNs share many parameters across all output states, and can therefore make more efficient use of the training data. GMMs often use diagonal covariance matrices to limit the number of parameters, and therefore assume that the input features are decorrelated across the input dimensions. It has often been observed that NN acoustic models outperform GMM acoustic models [12]. Considering the differences between these two model types, it is interesting to ask whether it is possible for them to learn from each other.

2.3. Input context

The HMM makes the assumption that the current observation is conditionally independent of all other states and observations, when given the current state. This assumption allows for efficient training and decoding, using methods such as the Viterbi algorithm [13]. However, this assumption places limitations on what the model can capture. To alleviate the impact of this assumption, hybrid models often use a context window of features as its input. NN architectures such as the Time Delay Neural Network (TDNN) [14] and Long Short-Term Memory (LSTM) [15] can utilise larger input context windows, while limiting the increase in the number of model parameters. However, these can be more computationally expensive to use for recognition. The forward computation through an LSTM is difficult to parallelise, due to the recurrent nature of the model, and must therefore process the inputs sequentially. Work in [5] trains a feed-forward DNN student to emulate a recurrent teacher, to allow for faster recognition. However, the DNN student has a narrower input context, and therefore has less information about the input to leverage upon. It is interesting to consider how the reduced input context of the student may affect its ability to learn from the teacher.

2.4. Feature representation

When teacher-student learning is used for domain adaptation, the teacher and student are often trained on two separate but aligned sets of input features. For example, when applied to far-field speech recognition [3], the teacher is trained on inputs from a close-talking microphone, while the student is trained on far-field inputs. It is hoped that the far-field student can perform similarly to the close-talking teacher. Another example of applying teacher-student learning for domain adaptation is in speaker adaptation [16]. Here, the teacher can be trained with per-speaker Constrained Maximum Likelihood Linear Regression (CMLLR) transforms [17] applied to the features, while the student uses features without CMLLR transforms. Transcriptions are required to train the CMLLR transforms for each speaker, and therefore when not using teacher-student learning, a two-pass recognition scheme is often used to obtain the CMLLR transforms for unseen speakers. This is computationally expensive and the reliability of the CMLLR transforms depends on the accuracy of the first recognition pass. When using teacher-student learning for speaker adaptation, the student uses features without CMLLR transforms, and therefore only a single recognition pass is needed. However, the teacher uses a different CMLLR transform on the inputs of each speaker, while the student is required to emulate this behaviour using the same inputs for all speakers.

Other than for the purpose of domain adaptation, it may also be beneficial to train a student toward teachers that use different input features, to, for example, gain from the diversity of an ensemble that uses diverse features. Such an ensemble can leverage upon the wide variety of possible feature representations that have been proposed for ASR, such as Mel-scale filterbank (FBK), Mel Frequency Cepstral Coefficients (MFCC) [18], and Perceptual Linear Predictive (PLP) [19] features. When using either different input sources or input feature representations, the different inputs may contain different information and may place different emphases on different aspects of the input information. Combining together models that use different features in an ensemble may benefit from the diversity of possible behaviours that can emerge from these differences.

However, the student is only privy to the information expressed in its own features, which may be different from that expressed in the features of the teacher. The experiments in this paper assess how this different input information can affect the student's ability to learn.

3. SEQUENCE-LEVEL TEACHER-STUDENT LEARNING

The aim of teacher-student learning is for the student to develop a behaviour that is similar to that of the teacher. One approach to train the student is to minimise the KL-divergence between the per-frame state posteriors of the teacher and student [2],

$$\mathcal{F}_{\text{frm}}\left(\boldsymbol{\Theta}\right) = -\sum_{t=1}^{T}\sum_{s_t} P\left(s_t|\mathbf{o}_t, \boldsymbol{\Phi}\right) \log P\left(s_t|\mathbf{o}_t, \boldsymbol{\Theta}\right), \quad (1)$$

where s_t are the states clusters, t is the frame index, T is the total number of frames, $\mathbf{o}_t$ are the input features, $\boldsymbol{\Phi}$ is the teacher model, and $\boldsymbol{\Theta}$ is the student model. However, this criterion only propagates information about the per-frame posteriors of the teacher. ASR is a sequence-to-sequence classification task, and this criterion may not effectively communicate information about the teacher's sequence-level behaviour. Furthermore, sequence-level training often yields a better performance than frame-level training when training toward the reference transcriptions [20].

Teacher-student learning can be generalised to the sequence-level, by minimising the KL-divergence between state sequence posteriors [4],

$$\mathcal{F}_{\text{seq}}\left(\boldsymbol{\Theta}\right) = -\sum_{\mathbf{s}_{1:T}} P\left(\mathbf{s}_{1:T}|\mathbf{O}_{1:T}, \boldsymbol{\Phi}\right) \log P\left(\mathbf{s}_{1:T}|\mathbf{O}_{1:T}, \boldsymbol{\Theta}\right), \quad (2)$$

where $\mathbf{s}_{1:T}$ are the state sequence hypotheses and $\mathbf{O}_{1:T}$ is the input feature sequence. Here, a sum over utterances is omitted for brevity. The gradient with respect to the student's observation log-likelihoods is

$$\frac{\partial \mathcal{F}_{\text{seq}}\left(\boldsymbol{\Theta}\right)}{\partial \log p\left(\mathbf{o}_t|s_t, \boldsymbol{\Theta}\right)} = \gamma\left[P\left(s_t|\mathbf{O}_{1:T}, \boldsymbol{\Theta}\right) - P\left(s_t|\mathbf{O}_{1:T}, \boldsymbol{\Phi}\right)\right], \quad (3)$$

where γ is the acoustic scaling factor that is often included to adjust the balance between the dynamic ranges of the acoustic and language models. Both $P\left(s_t|\mathbf{O}_{1:T}, \boldsymbol{\Theta}\right)$ and $P\left(s_t|\mathbf{O}_{1:T}, \boldsymbol{\Phi}\right)$ can be computed using a forward-backward operation, over the student's and teacher's denominator lattices respectively. It is possible to compute this gradient using a lattice-free framework [21].

4. TEACHER-STUDENT LEARNING WITH A GMM

Two commonly used acoustic model types are based on NNs and GMMs. This paper explores the possibility of propagating informa-

tion between them. The GMM acoustic model computes the per-frame observation likelihoods as

$$p\left(\mathbf{o}_t|s_t,\boldsymbol{\Theta}\right)=\sum_{k=1}^{K_{s_t}}\lambda_{ks_t}\mathcal{N}\left(\mathbf{o}_t|\boldsymbol{\mu}_{ks_t},\boldsymbol{\Sigma}_{ks_t}\right), \tag{4}$$

where k is the mixture component index, K_s is the number of mixture components for state s, $\mathcal{N}$ is a Gaussian density function, and the model parameters, $\boldsymbol{\Theta}$, consist of the means, $\boldsymbol{\mu}_{ks}$, covariances, $\boldsymbol{\Sigma}_{ks}$, and mixture weights, λ_{ks}. The mixture weights satisfy $\sum_k \lambda_{ks}=1$ and $\lambda_{ks}\geq 0$. With sufficient mixture components, the GMM can potentially model any probability density function shape of the observations for each state.

A NN acoustic model is often designed to compute state posteriors, $P\left(s_t|\mathbf{o}_t,\boldsymbol{\Theta}\right)$. The NN performs multiple levels of non-linear and possibly recurrent operations on the observations, with a final softmax performed on the output to compute a normalised state posterior distribution. Un-normalised observation likelihoods can then be obtained as

$$p\left(\mathbf{o}_t|s_t,\boldsymbol{\Theta}\right)\propto\frac{P\left(s_t|\mathbf{o}_t,\boldsymbol{\Theta}\right)}{P\left(s_t\right)}. \tag{5}$$

The GMM and NN compute the observation likelihoods differently, and therefore may yield highly diverse behaviours. Considering these differences, it is interesting to question whether it is possible for these acoustic models to learn from each other using teacher-student learning. NN acoustic models can readily be used with frame-level teacher-student learning of (1). However, without state posteriors, it is not trivial to use GMM acoustic models with (1).

Sequence-level teacher-student learning with (2) does not require per-frame state posteriors. All that is required is that it must be possible to compute the state sequence posteriors, $P\left(\mathbf{s}_{1:T}|\mathbf{O}_{1:T},\boldsymbol{\Theta}\right)$, from the models. It is therefore possible to use sequence-level teacher-student learning with a GMM acoustic model. This can be used to, for example, train a GMM student to learn from a NN teacher.

The GMM acoustic model is often trained using the Baum-Welch [22] or Extended Baum-Welch (EBW) [23] algorithms, which are instances of the expectation-maximisation algorithm. It is also possible to train a GMM with gradient descent [24, 25]. However, doing so sacrifices the guarantee that the training criterion will not worsen at each iteration. Diagonal covariance matrices are used in this paper. When using gradient descent, it needs to be enforced that the mixture component weights satisfy $\sum_k \lambda_{ks}=1$ and $\lambda_{ks}\geq 0$, and that the diagonal variances satisfy $\sigma_{iks}\geq 0$, where i is the input dimension index. These can be enforced by re-parameterising the GMM with $\widetilde{\lambda}$ and $\widetilde{\sigma}$, such that

$$\lambda_{ks}=\frac{\exp\left[\widetilde{\lambda}_{ks}\right]}{\sum_{k'=1}^{K_s}\exp\left[\widetilde{\lambda}_{k's}\right]} \tag{6}$$

and

$$\sigma_{iks}=\exp\left[\widetilde{\sigma}_{iks}\right]. \tag{7}$$

Note that in the notation used here, σ represents the variance, not the standard deviation. The trainable parameters of the GMM are then $\boldsymbol{\Theta}=\left\{\boldsymbol{\mu}_{ks},\widetilde{\boldsymbol{\sigma}}_{ks},\widetilde{\lambda}_{ks}\quad\forall k,s\right\}$. The derivatives of the per-frame observation log-likelihoods with respects to the parameters are [24]

$$\frac{\partial\log p\left(\mathbf{o}|s,\boldsymbol{\Theta}\right)}{\partial\mu_{iks}}=\frac{\lambda_{ks}\mathcal{N}\left(\mathbf{o}|\boldsymbol{\mu}_{ks},\boldsymbol{\Sigma}_{ks}\right)}{p\left(\mathbf{o}|s,\boldsymbol{\Theta}\right)}\frac{o_i-\mu_{iks}}{\sigma_{iks}} \tag{8}$$

$$\frac{\partial\log p\left(\mathbf{o}|s,\boldsymbol{\Theta}\right)}{\partial\widetilde{\sigma}_{iks}}=\frac{\lambda_{ks}\mathcal{N}\left(\mathbf{o}|\boldsymbol{\mu}_{ks},\boldsymbol{\Sigma}_{ks}\right)}{p\left(\mathbf{o}|s,\boldsymbol{\Theta}\right)}\frac{\left(o_i-\mu_{iks}\right)^2-\sigma_{iks}}{2\sigma_{iks}} \tag{9}$$

$$\frac{\partial\log p\left(\mathbf{o}|s,\boldsymbol{\Theta}\right)}{\partial\widetilde{\lambda}_{ks}}=\lambda_{ks}\left[\frac{\mathcal{N}\left(\mathbf{o}|\boldsymbol{\mu}_{ks},\boldsymbol{\Sigma}_{ks}\right)}{p\left(\mathbf{o}|s,\boldsymbol{\Theta}\right)}-1\right]. \tag{10}$$

Here, the frame index, t, has been omitted for easier readability. These derivatives can be combine together with (3) using the chain rule, to compute the parameter gradients to train a GMM student.

5. EXPERIMENTS

The experiments aim to assess the ability of the student to learn, when differences exist between the designs of the student and teacher. These were implemented using the Kaldi speech recognition toolkit [26]. Two datasets were used. The AMI meeting transcription task [27] comprises spontaneous speech from multiple speakers in role-play meeting scenarios. The *full corpus ASR partition* was used, consisting of an 81 hours training set and a 9 hours *eval* set. The Individual Headset Microphone (IHM) audio recordings were used. The 2017 Multi-Genre Broadcast (MGB-3) English task [28] comprises audio recordings from television programs of a variety of genres. Lightly supervised decoding and selection [29] was used to extract a training set with 275 hours of data, out of the full 375 hours of available audio data. The 5.5 hours *dev17b* test set was used, and was divided into segments using a DNN-based segmenter [30] that was trained on the MGB-3 data.

The NN models used in the experiments were trained with a lattice-free implementation of either the Maximum Mutual Information (MMI) criterion [31] or the sequence-level teacher-student learning criterion of (2), beginning from a random parameter initialisation. Minimum Bayes' Risk (MBR) decoding [32] was used to perform recognition. In MGB-3, decoding was done using a trigram language model, trained on the MGB-3 subtitle data. In AMI-IHM, decoding lattices were first generated using a trigram language model, then rescored using a 4-gram language model, both trained on a combination of the AMI training set and Fisher English training part 1 (LDC2004T19) transcriptions.

5.1. Different model types, complexities, and input contexts

The first experiment assesses the ability of a student to learn, when its model type, model complexity, and input context differ from that of the teacher. Different models were first trained using the lattice-free MMI criterion, to be used as the teachers. The diagonal-covariance GMM for AMI-IHM had 20 mixture components per state and used 13-dimensional MFCC features, to allow a reasonable match with the diagonal covariance. Only 13 dimensions were used to limit the number of parameters in the GMM. First and second temporal derivatives were appended to the GMM inputs. The lattice-free GMM was trained using gradient descent, to provide a comparable baseline for the students, which were also trained with gradient descent. NN acoustic models used 40-dimensional Mel-scaled FBK features as inputs. Two feed-forward DNNs were used, a large DNN had 17 layers, while a smaller model with 4 layers is referred to as $\text{DNN}_{\text{small}}$. These DNNs used a symmetric input context of 9 spliced frames, which matches the temporal context of the GMM's temporal

derivatives. A 17 layer TDNN was also used, with a symmetric total input context of 69 frames. Both the DNN and TDNN layers used rectified linear unit activations, had residual connections, and were factorised (these are often referred to as TDNN-F in the literature [33], but the F is omitted here for simplicity). Finally, an acoustic model with interleaved TDNN and uni-directional LSTM layers was also used in MGB-3, and is referred to as TDNN-LSTM. The TDNN-LSTM has a potentially infinite backward context, and used a forward context of 23 frames (excluding the current frame).

The performances of these models with lattice-free MMI training for AMI-IHM and MGB-3 are shown in the right side columns of Tables 1 and 2 respectively. As a reference, a GMM that was trained with an EBW lattice-based implementation of the MMI criterion has a WER of 37.7% in AMI-IHM. Several differences exist between this lattice-based GMM and the lattice-free GMM in Table 1. The lattice-based GMM used a triphone decision tree with 4000 leaves, while the lattice-free GMM used a left-biphone decision tree with 2000 leaves for efficient training. Also, the lattice-based GMM used a frame shift of 10ms and a 3-state HMM with trained transition probabilities, while the lattice-free GMM used a frame shift of 30ms and a 2-state HMM with uniform transition probabilities, again for efficient training. The lattice-based GMM was first trained with a maximum likelihood criterion, then fine-tuned with MMI, while the lattice-free GMM was trained toward the MMI criterion, beginning from a random parameter initialisation. The lattice-based GMM used a variable number of mixture components per state, and the number of mixture components was grown by splitting the most likely components at regular training iteration intervals. As opposed to this, the lattice-free GMM implementation used here fixed all states to have the same number of mixture components, to simplify the GPU-based feed-forward and back-propagation. These differences may account for the performance degradation of the lattice-free GMM. Work in [24] suggests several methods to improve gradient descent training of a GMM. Out of these, it was found that L2 regularisation of the GMM parameters greatly improved the performance.

Table 1. Student WER (%) when learning from teachers with different acoustic models and input contexts, using sequence-level Teacher-Student (TS) learning, on AMI-IHM

	Sequence TS toward			lattice-free
Student	GMM	DNN	TDNN	MMI
GMM	46.6	41.7	45.4	45.0
DNN_{small}	42.2	28.6	29.1	28.8
DNN	42.0	27.8	27.9	27.7
TDNN	41.3	26.6	22.1	22.9

The NN models are able to outperform both the lattice-based and lattice-free GMMs. The results also show that increasing the acoustic model's complexity and input context can yield an improved performance.

Students with a variety of acoustic model types were trained toward these different models as teachers, using sequence-level teacher-student learning. The performances of these students for AMI-IHM and MGB-3 are shown in the left three columns of Tables 1 and 2 respectively. From these results, it can be seen that several of the students outperform their teacher. This is especially so when the student uses a more complex model or has a wider input context than the teacher. During teacher-student learning, the student is trained to produce similar state sequence posteriors as the teacher on the training set. However, during recognition, the performance of the student is measured based on its word sequence hypotheses on an unseen test set. As such, the teacher does not strictly represent a lower bound of the WER performance for the student, and it is possible that the student may generalise to unseen data better than the teacher. In AMI-IHM, the TDNN student outperforms its TDNN teacher. The TDNN teacher and TDNN student have WERs of 8.3 and 8.8% respectively when measured on the training set, and WERs of 22.9 and 22.1% when measured on the unseen *eval* set from Table 1. This suggests that it is possible for the student to generalise better to unseen data than the teacher.

Table 2. Student WER (%) when learning from teachers with different acoustic models and input contexts, on MGB-3

	Sequence TS toward			lattice-free
Student	DNN	TDNN	TDNN-LSTM	MMI
DNN_{small}	31.7	31.4	30.1	32.5
DNN	30.4	29.1	28.1	29.5
TDNN	29.1	22.5	21.9	22.6
TDNN-LSTM	29.2	22.4	21.3	21.4

When both the teacher and student use the same TDNN acoustic model, is surprising that the student develops better generalisation behaviour, considering that the global optimum for this teacher-student learning optimisation problem is for the student's model parameters to be equal to those of the teacher. One possible explanation for the better generalisation ability of the student is that the student may have converged to a local optimum that has different parameters from the teacher, suggesting that teacher-student learning is sensitive to the initialisation of the student. Another possibility is that the exponentially decaying learning rate schedule and gradient-based optimisation that were used may not have allowed the student to reach an optimum, thereby yielding a regularisation effect, similar to early stopping.

Comparing the model type, it can be seen in AMI-IHM that a GMM can learn from a NN teacher, yielding a better performance than lattice-free MMI training of the GMM and a GMM student learning from a GMM teacher. It is therefore possible to effectively propagate information across these different acoustic model types by using sequence-level teacher-student learning.

Next, a student can be trained toward a larger teacher. As a reference, a DNN_{small} student learning from a DNN_{small} teacher has a WER of 29.1 and 33.3% for AMI-IHM and MGB-3 respectively. Learning from the larger DNN teacher improves the student performance. The DNN_{small} students of the DNN teachers also perform better than DNN_{small} models trained with lattice-free MMI, corroborating the results in [2]. The results suggest that a smaller DNN_{small} student can benefit by learning from the larger DNN teacher.

Finally, a student can be trained toward a teacher with a different input context. From the AMI-IHM results in Table 1, the GMM and DNN_{small} students learn better from a DNN teacher than from a TDNN teacher, even though the TDNN teacher performs better than the DNN teacher. The DNN teacher has the same input context as these students. In MGB-3, the DNN_{small} students perform better when learning from the TDNN and TDNN-LSTM teachers, rather than the DNN teacher. This may suggest that using more training data may help the student to learn to better accommodate for its reduced input context. However, the TDNN and TDNN-LSTM teachers have much larger performance improvements over the lattice-free MMI DNN_{small} model, compared to the performance difference between DNN_{small} students trained toward the TDNN or TDNN-LSTM teachers, and toward the DNN teacher. The trend is the same for the

larger DNN student in MGB-3. These results suggest that it may be difficult for a student to learn from a teacher with a wider input context. The student may not have access to the input information that would allow it to effectively emulate the teacher. The results also show that when the student has a wider input context than the teacher, then learning is effective.

5.2. Different input features

The previous experiment investigated teacher-student learning between different acoustic model types, model complexities, and input contexts. The next experiment assesses the ability of the student to learn when its input feature representation differs from that used by the teacher. All acoustic models used in this experiment were TDNNs. In AMI-IHM, the same 40-dimensional FBK_{40} and 13-dimensional $MFCC_{13}$ features from the previous experiment were used, without any temporal derivatives. In this experiment, the feature dimension is explicitly written in the subscript for clarity. The $MFCC_{13}$ features were computed by taking a linear Discrete Cosine Transform (DCT) of the FBK_{40} features, then retraining only the first 13 dimensions. This truncation may result in information loss in the features. As a comparison, 40-dimensional $MFCC_{40}$ features were also used, without truncating the DCT output. The DCT is a full-rank linear transform, and therefore should not result in any information loss. In MGB-3, a comparison is made between FBK_{40} and 13-dimensional PLP_{13} features. The feature extraction pipelines for FBK and MFCC features are identical, differing only by a DCT [18]. As opposed to this, the extraction pipeline for PLP [19] is significantly different, and it may therefore express information differently. When a student learns from a single teacher that uses the same input features, there may not be much to gain. Therefore, a better teacher that used the same input features and the same input context was constructed by combining an ensemble of 4 models, each trained beginning from a different random parameter initialisation. The performances of these single and ensemble teachers are shown in Table 3. Ensemble combination was performed using MBR combination decoding [32]. The cross-WER (cWER) [34] provides an approximate measure of the ensemble diversity. This computes the word-level minimum edit distance between the hypotheses of two models, normalised by the hypothesis length, averaged over all pairs of models in the ensemble. A larger cWER indicates a wider diversity of hypotheses.

Table 3. Single models and random initialisation ensembles with FBK, MFCC, and PLP features

Dataset	Feature	Single WER (%) mean	Single WER (%) std dev	Ensemble WER (%)	Diversity cWER (%)
AMI-IHM	FBK_{40}	22.9	0.1	21.1	14.3
	$MFCC_{40}$	23.0	0.1	21.1	14.7
	$MFCC_{13}$	24.1	0.1	22.2	15.1
MGB-3	FBK_{40}	22.6	0.2	21.0	12.1
	PLP_{13}	24.9	0.2	23.2	13.8

The $MFCC_{40}$ model is able to perform comparably to the FBK_{40} model, while there are performance degradations for the $MFCC_{13}$ and PLP_{13} models, indicating the detrimental impact of information loss in the input features. It can also be seen that for all feature types, generating ensembles by simply training multiple models from different random parameter initialisations is able to yield diverse behaviours and combination gains over the respective single models. For each of the ensembles, the cWER is a significant fraction of the combined WER, indicating a wide diversity of behaviours among the constituent models.

Table 4. Student WER (%) when learning from teachers with FBK and MFCC features, on AMI-IHM

Model	Single teacher FBK_{40}	Single teacher $MFCC_{40}$	Single teacher $MFCC_{13}$	Ensemble Teacher FBK_{40}	Ensemble Teacher $MFCC_{40}$	Ensemble Teacher $MFCC_{13}$
Student						
FBK_{40}	22.1	22.1	22.8	21.5	21.6	22.2
$MFCC_{40}$	21.9	22.2	22.8	21.5	21.4	22.1
$MFCC_{13}$	23.0	23.1	23.2	22.6	22.4	22.6
Teacher	22.9	23.0	24.1	21.1	21.1	22.2

Students were trained toward the teachers with sequence-level teacher-student learning, using the variety of input features. The results are shown in Tables 4 and 5 for AMI-IHM and MGB-3 respectively.

Table 5. Student WER (%) when learning from teachers with FBK and PLP features, on MGB-3

Model	Single teacher FBK_{40}	Single teacher PLP_{13}	Ensemble teacher FBK_{40}	Ensemble teacher PLP_{13}
Student				
FBK_{40}	22.5	23.9	22.0	23.4
PLP_{13}	24.2	24.8	23.7	24.2
Teacher	22.6	24.9	21.0	23.2

Similarly to the results in Table 1, several of the AMI-IHM students in Table 4 also perform better than their single model teachers. However, for both datasets, the students are not able to perform better than the ensemble teachers. In AMI-IHM, the FBK_{40} and $MFCC_{40}$ students show comparable performances when learning from either the FBK_{40} or $MFCC_{40}$ teachers. This suggests that the NN acoustic model is able to accommodate for the difference in the input representation caused by the DCT. The $MFCC_{13}$ student is not able to gain from the better performances of either the FBK_{40} or $MFCC_{40}$ teachers. In MGB-3, the PLP_{13} student yields a better performance when learning from a FBK_{40} teacher, over learning from a PLP_{13} teacher. This may suggest that using more training data may allow the student to better learn to accommodate for deficiencies in its input representation. However, the improvement in the student, when learning from the FBK_{40} teacher rather than the PLP_{13} teacher, is much smaller than the performance improvement that the FBK_{40} teacher has over the PLP_{13} teacher. These results suggest that it may be difficult for the student to learn from the teachers about how to correct for the information loss in its input features.

Table 6. Single models and random initialisation ensembles with MFCC and MFCC-SA features, on AMI-IHM

Feature	Single WER (%) mean	Single WER (%) std dev	Ensemble WER (%)	Diversity cWER (%)
MFCC	24.1	0.1	22.2	15.1
MFCC-SA	22.3	0.2	20.8	13.4

In a speaker adaptation scenario, following the example of [16], a teacher can be trained with CMLLR transforms applied to the input features, while the student is trained without the CMLLR transforms. At the teacher's input, a different CMLLR transform is ap-

plied to the features from each speaker. In contrast, the FBK_{40} student and $MFCC_{40}$ teacher in Table 4 are differentiated by a fixed linear transform. The student in the speaker adaptation scenario is therefore tasked with learning to behave as if speaker-specific transforms were used, even though the student has no speaker-specific transforms. The final experiment investigates the impact of these speaker-specific transforms on the student's ability to learn. This experiment is performed on the AMI-IHM dataset. Per-speaker CMLLR transforms were obtained using an initial GMM model, trained with a lattice-based maximum likelihood criterion. These used 13-dimensional MFCC features as input. In this experiment, the feature dimension subscript is omitted for simplicity. The CMLLR transforms were applied to the MFCC features, and the result is referred to as MFCC-SA features, where SA stands for Speaker Adapted. For speakers in the *eval* set, the first-pass recognition hypotheses used to train the CMLLR transforms were obtained using a maximum likelihood GMM model that used MFCC features without the CMLLR transforms. Single lattice-free MMI TDNN models were trained on MFCC and MFCC-SA features. Ensembles were also constructed to provide better teachers for each feature type. These were again generated by training 4 models, beginning from different random parameter initialisations. The performances of these teachers are shown in Table 6.

Table 7. Student WER (%) when learning from teachers with MFCC and MFCC-SA features, on AMI-IHM

	Single teacher		Ensemble teacher	
Model	MFCC	MFCC-SA	MFCC	MFCC-SA
Student				
MFCC	23.2	23.4	22.6	22.6
Teacher	24.1	22.3	22.2	20.8

Students that used MFCC features were trained toward each of the teachers, and the results are shown in Table 7. The MFCC student learning from the MFCC-SA teacher performs better than an MFCC model that is trained with lattice-free MMI. This agrees with previous results in [16], and shows that teacher-student learning can be used for speaker adaptation. However, this student does not perform better than an MFCC student learning from an MFCC teacher. As a reference, a student that uses MFCC-SA features as input and learns from the single MFCC-SA teacher has a WER of 22.0%, surpassing the teacher. The *eval* set CMLLR transforms for this student were the same as those used for the MFCC-SA models in Table 6, computed with first-pass hypotheses from an initial GMM. The MFCC student performs worse than the MFCC-SA student when learning from the MFCC-SA teacher. As opposed to this, the FBK_{40} and $MFCC_{40}$ students perform comparably when learning from either of the FBK_{40} or $MFCC_{40}$ teachers in Table 4. These results suggest that it is in fact difficult for the MFCC student to emulate the per-speaker CMLLR transforms of the teacher.

6. CONCLUSION

This paper has studied the ability of a student to learn from a teacher with different acoustic model types, acoustic model complexities, input contexts, and input feature representations. The results suggest that the student can effectively learn from a teacher with the same input context and features, but with a larger model complexity. However, the results also suggest that the student struggles to effectively learn when it has a narrower input context, has information loss in its input features, or if only the teacher uses differing transforms for the inputs of each speaker. Using more training data may slightly improve the student's ability to learn to overcome the deficiencies in its input context and features. It may therefore be important to carefully consider the design of the student for each application. This paper has also presented a first attempt to propagate information from a NN teacher to a GMM student, and the results show that this information can improve the GMM student performance.

7. REFERENCES

[1] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *KDD*, Philadelphia, USA, Aug 2006, pp. 535–541.

[2] J. Li, R. Zhao, J-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Interspeech*, Singapore, Sep 2014, pp. 1910–1914.

[3] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," in *Interspeech*, Stockholm, Sweden, Aug 2017, pp. 2386–2390.

[4] J. H. M. Wong and M. J. F. Gales, "Sequence student-teacher training of deep neural networks," in *Interspeech*, San Francisco, USA, Sep 2016, pp. 2761–2765.

[5] W. Chan, N. R. Ke, and I. Lane, "Transferring knowledge from a RNN to a DNN," in *Interspeech*, Dresden, Germany, Sep 2015, pp. 3264–3268.

[6] J. H. M. Wong and M. J. F. Gales, "Student-teacher training with diverse decision tree ensembles," in *Interspeech*, Stockholm, Sweden, Aug 2017, pp. 117–121.

[7] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*, Kluwer Academic Publishers, 1994.

[8] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *ICASSP*, Shanghai, China, Mar 2016, pp. 4960–4964.

[9] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML Representation Learning Workshop*, Edinburgh, UK, Jul 2012.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, Pittsburgh, USA, Jun 2006.

[11] B.-H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Technical Journal*, vol. 64, no. 6, pp. 1235–1249, Jul-Aug 1985.

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

[13] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, Apr 1967.

[14] A. Waibel, T. Hanazawa, G. E. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, Mar 1989.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

[16] N. M. Joy, S. R. Kothinti, S. Umesh, and B. Abraham, "Generalized distillation framework for speaker normalization," in *Interspeech*, Stockholm, Sweden, Aug 2017, pp. 739–743.

[17] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, Apr 1998.

[18] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed., pp. 374–388. Academic Press, 1976.

[19] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr 1990.

[20] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *ICASSP*, Taipei, Apr 2009, pp. 3761–3764.

[21] N. Kanda, Y. Fujita, and K. Nagamatsu, "Investigation of lattice-free maximum mutual information-based acoustic models with sequence-level Kullback-Leibler divergence," in *ASRU*, Okinawa, Japan, Dec 2017, pp. 69–76.

[22] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," in *Symposium on Inequalities*, Los Angeles, USA, 1972, pp. 1–8.

[23] P. S. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "A generalization of the Baum algorithm to rational objective functions," in *ICASSP*, Glasgow, UK, May 1989, pp. 631–634.

[24] C. Zhang and P. C. Woodlang, "Joint optimisation of tandem systems using Gaussian mixture density neural network discriminative sequence training," in *ICASSP*, New Orleans, USA, Mar 2017, pp. 2015–5019.

[25] Y. Wang, C. Zhang, M. J. F. Gales, and P. C. Woodland, "Speaker adaptation and adaptive training for jointly optimised tandem systems," in *Interspeech*, Hydrabad, India, Sep 2018, pp. 872–876.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *ASRU*, Hawaii, USA, Dec 2011.

[27] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: a pre-announcement," in *MLMI*, Edinburgh, UK, July 2005, pp. 28–39.

[28] P. Bell, "MGB challenge," May 2017, http://www.mgb-challenge.org/english.html.

[29] P. Lanchantin, M. J. F. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P. C. Woodland, and C. Zhang, "Selection of multi-genre broadcast data for the training of automatic speech recognition systems," in *Interspeech*, San Francisco, USA, Sep 2016, pp. 3057–3061.

[30] L. Wang, C. Zhang, P. C. Woodland, M. J. F. Gales, P. Karanasou, P. Lanchantin, X. Liu, and Y. Qian, "Improved DNN-based segmentation for multi-genre broadcast audio," in *ICASSP*, Shanghai, China, Mar 2016, pp. 5700–5704.

[31] 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 *Interspeech*, San Francisco, USA, Sep 2016, pp. 2751–2755.

[32] 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, Oct 2011.

[33] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech*, Hydrabad, India, Sep 2018, pp. 3743–3747.

[34] J. H. M. Wong and M. J. F. Gales, "Multi-task ensembles with teacher-student training," in *ASRU*, Okinawa, Japan, Dec 2017, pp. 84–90.