General sequence teacher-student learning

Jeremy H. M. Wong, *Student Member, IEEE,* Mark J. F. Gales, *Fellow, IEEE,* and Yu Wang, *Member, IEEE*

***Abstract*—In automatic speech recognition, performance gains can often be obtained by combining an ensemble of multiple models. However, this can be computationally expensive when performing recognition. Teacher-student learning alleviates this cost by training a single student model to emulate the combined ensemble behaviour. Only this student needs to be used for recognition. Previously investigated teacher-student criteria often limit the forms of diversity allowed in the ensemble, and only propagate information from the teachers to the student at the frame level. This paper addresses both of these issues by examining teacher-student learning within a sequence-level framework, and assessing the flexibility that these approaches offer. Various sequence-level teacher-student criteria are examined in this work, to propagate sequence posterior information. A training criterion based on the KL-divergence between context-dependent state sequence posteriors is proposed that allows for a diversity of state cluster sets to be present in the ensemble. This criterion is shown to be an upper bound to a more general KL-divergence between word sequence posteriors, which places even fewer restrictions on the ensemble diversity, but whose gradient can be expensive to compute. These methods are evaluated on the AMI meeting transcription and MGB-3 television broadcast audio tasks.**

***Index Terms*—Automatic speech recognition, ensemble, lattice-free, random forest, teacher-student**

I. Introduction

Teacher-student learning [1] is a framework that can be used to transfer knowledge between models. In Automatic Speech Recognition (ASR), this has found application in compressing a large model [2] or an ensemble of models [1], and in domain adaptation [3]. The standard teacher-student learning method trains a student by minimising the KL-divergence between per-frame state cluster posteriors [2], [4]. However, frame-level training does not consider the sequential nature of speech data and constrains all models to use the same set of state clusters. This paper proposes a generalisation of the teacher-student learning framework to overcome both of these limitations.

One common application of teacher-student learning is ensemble compression. In ASR, performance gains can often be obtained by combining an ensemble of multiple models together, over using just a single model [5], [6]. A review of ensemble methods is presented in Section II. The ensemble performance depends on both the accuracy of the constituent models and the diversity between the model behaviours [7]. A rich ensemble can be generated by allowing for diversity in the acoustic model, sub-word units, context-dependency, and state clusters, to name a few. It may be possible to use these different forms of diversity together to obtain a more diverse ensemble. In previous investigations of ensemble methods, it has been common to only capture diversity within the acoustic model, either in the parameters [8] or the topology [9], while constraining the sets of sub-word units and state clusters to be the same across all members of the ensemble. This may limit the diversity that can be captured within the ensemble, and therefore also limit the combination gains. Work in [10] has introduced a new degree of diversity, by allowing the sets of state clusters to differ between models. Work in [11] has investigated using different sets of sub-word units.

Although an ensemble may perform well, it can be computationally expensive to use for recognition. Teacher-student learning [1] is one method that aims to alleviate this computational demand, by training a single student model to emulate the combined ensemble. During recognition, only this student needs to be used. A commonly used method to train the student is to minimize the KL-divergence between the teachers' and student's frame-level state cluster posteriors [2]. This method is discussed in Section III. However, this criterion limits the forms of diversity that the ensemble is allowed to have, by constraining all models to use the same set of state clusters. This in turn forces all models to also use the same set of sub-word units, context-dependency, and Hidden Markov Model (HMM) topology. Teacher-student learning can be generalised to allow for different sets of state clusters between models, by instead minimising the KL-divergence between logical context-dependent state posteriors [12]. This allows the ensemble to have a diversity of state cluster sets.

However, the per-frame posteriors from the teachers may not fully capture the sequential nature of speech data. For standard ASR training, sequence training has often been found to yield a better performance than frame-level training [13]. As is discussed in Section IV, teacher-student learning can be extended to the sequence level [8], allowing information about the sequence-level behaviours of the teachers to be propagated to the student. This paper considers minimising the KL-divergence between word sequence posteriors as one possible criterion. This criterion places few constraints on the allowed forms of diversity in the ensemble. However, it is shown that the gradient of this criterion can be expensive to compute. To address this, the student can be trained by minimising the KL-divergence between lattice arc sequence posteriors. Previous work has investigated using such a criterion, where the arcs are marked with state clusters [8]. The gradient of this criterion can be computed efficiently, but again constrains all models to use the same set of state clusters, and thereby limits the ensemble diversity. It is also possible to implement this criterion in a lattice-free framework [14].

This paper proposes a generalisation of teacher-student learning that allows different state cluster sets between models, while using a sequence-level criterion. This extends the work in [12] to the sequence-level and generalises the work in [8] to allow for different state cluster sets. To enable this, lattice

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

arcs are marked with logical context-dependent states instead of state clusters, and the student can be trained by minimising the KL-divergence between logical context-dependent state sequence posteriors. The proposed training method can be implemented in both lattice-based and lattice-free frameworks. In this paper, a relationship is also drawn between the lattice arc sequence and word sequence posterior KL-divergences, showing the former to be an upper bound to the latter.

II. Ensemble methods

Performance gains can often be obtained by combining together an ensemble of multiple models [5]. The combined ensemble performance depends on both the individual model performances and the diversity between the model behaviours [7]. There are many ways that models can be made to behave differently, such as by using different sets of model parameters and topologies, feature representations, sub-word units, and state clusters. Using more forms of diversity may allow for a richer ensemble. This paper considers using diverse acoustic models parameters and state cluster sets.

A. Parameter diversity

Often, when generating an ensemble of models, only limited forms of diversity are introduced. One simple approach is to use different sets of acoustic model parameters, while constraining the acoustic model topology and set of state clusters to be the same. One method of generating this form of ensemble is to train multiple models, each beginning from a different random parameter initialisation [8]. However, this can be computationally expensive to train. A cheaper method of obtaining a diversity of model parameters is to use the intermediate model iterations from within a single run of training as members of the ensemble [15].

However, ensembles of these forms may be limited in their diversity. To allow for added diversity, previous works have explored combining different acoustic model topologies [9], sets of state clusters [10], and sets of sub-word [11].

B. State cluster diversity

This paper considers ensembles where the set of state clusters is allowed to vary between models. In ASR it has been found that the acoustic realisations of phones are strongly influenced by their surrounding contexts, leading to the use of context-dependent phones [16]. However, there are often too many context-dependent states to individually model robustly. As such, similar logical context-dependent states can be clustered together into state clusters, with their observation likelihoods tied. One commonly used method to perform state clustering is through a phonetic decision tree [16], $\mathcal{T}$,

$$s_c = \mathcal{T}(c), \tag{1}$$

which maps from a logical context-dependent state, c, to a state cluster, s. It is computationally intractable to find a globally optimal decision tree [17]. Trees are therefore often built by choosing the "greedy" split, with the largest local increase in likelihood, at each training iteration from a set of phonetically motivated questions [16]. However, this is not guaranteed to produce a result that is optimal over the whole tree.

When using hybrid Neural Network (NN)-HMM models, the set of state clusters defines the classes that the NN acoustic model discriminates between. An ensemble can be constructed by associating a different set of state clusters with each model [18]. Each member of the ensemble then learns to discriminate between different sets of state clusters. This may encourage diverse behaviours between the ensemble members. This form of ensemble diversity has been found to be particularly effective, especially when the quantity of training data is limited [12].

Multiple sets of state clusters can be obtained by using the random forest method [19]. In this method, multiple decision trees are generated by uniformly sampling from the n-best splits at each training iteration, instead of choosing the greedy split. Although there are other methods that explicitly train multiple decision trees to be different [20], the simple random forest method is used in this paper.

C. Ensemble combination

An ensemble can be combined at the hypothesis level [5], [6]. One such method is Minimum Bayes' Risk (MBR) combination decoding [6],

$$\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}'} \sum_{\boldsymbol{\omega}} \mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\omega}') P\left(\boldsymbol{\omega} \middle| \mathbf{O}, \widehat{\boldsymbol{\Phi}}\right), \tag{2}$$

where $\boldsymbol{\omega}$ are the word sequence hypotheses, $\mathbf{O}$ is the observation sequence, $\widehat{\boldsymbol{\Phi}}$ represents the ensemble, and $\mathcal{L}$ is the word-level minimum edit distance. The combined hypothesis posteriors can be computed as

$$P\left(\boldsymbol{\omega} \middle| \mathbf{O}, \widehat{\boldsymbol{\Phi}}\right) = \sum_{m=1}^{M} \lambda_m P(\boldsymbol{\omega}|\mathbf{O}, \boldsymbol{\Phi}^m), \tag{3}$$

where $\boldsymbol{\Phi}^m$ are the parameters of the mth model, M is the ensemble size, and λ_m are the interpolation weights that satisfy $0 \leq \lambda_m \leq 1$ and $\sum_m \lambda_m = 1$. However, this requires a separate decoding run for each model, and therefore has a computational cost that scales linearly with the ensemble size when performing recognition. This cost can be reduced by performing frame-level combination [21], as only one decoding run is required for the whole ensemble. However, data still needs to be fed through each separate acoustic model. One possible method to further reduce the computational cost is to instead perform combination over the model parameters, resulting in a single "smoothed" model, $\overline{\boldsymbol{\Phi}}$. One such combination method is to linearly interpolate the model parameters [22],

$$\overline{\boldsymbol{\Phi}} = \sum_{m=1}^{M} \eta_m \boldsymbol{\Phi}^m, \tag{4}$$

where η_m are the interpolation weights. Only the smoothed model needs to be used during recognition. However, this form of parameter-level combination can only be applied when all models in the ensemble use the same topology and whose hidden representations are ordered similarly. These constraints may limit the diversity in an ensemble that can be used with

this combination method. Generating an ensemble out of the intermediate models from within a single run of training is one method that abides by these constraints [23].

III. Teacher-student learning

Teacher-student learning is a more general framework that can reduce the computational cost of using an ensemble for recognition, with fewer constraints on the allowed forms of ensemble diversity than parameter-level combination. Here, a single student model is trained to emulate the behaviour of the combined ensemble [1]. Only this single student then needs to be used for recognition.

In ASR, the student should ideally produce similar hypotheses to the combined ensemble. The standard approach trains the student by minimising the KL-divergence between the frame-level state cluster posteriors of the combined ensemble and the student [2],

$$\mathcal{F}_{\text{TS}}^{\text{state}}\left(\boldsymbol{\Theta}\right) = -\sum_{t=1}^{T}\sum_{s\in\mathcal{T}} P\left(s\middle|\mathbf{o}_t,\widehat{\boldsymbol{\Phi}}\right)\log P\left(s|\mathbf{o}_t,\boldsymbol{\Theta}\right), \quad (5)$$

where t is the frame index, T is the number of frames in an utterance, $\mathbf{o}_t$ are the observations, $\boldsymbol{\Theta}$ are the student model parameters, and $\mathcal{T}$ represents the set of state clusters defined by the decision tree. An implied summation over utterances is omitted in the criteria presented in this paper for brevity. This criterion propagates per-frame state cluster posterior information from the teachers to the student. The targets are the combined frame-level state cluster posteriors of the ensemble,

$$P\left(s\middle|\mathbf{o}_t,\widehat{\boldsymbol{\Phi}}\right) = \sum_{m=1}^{M}\lambda_m P\left(s|\mathbf{o}_t,\boldsymbol{\Phi}^m\right). \quad (6)$$

Although a sum combination is considered here, there are also other possible ways of combining the teacher posteriors [21]. A limitation of $\mathcal{F}_{\text{TS}}^{\text{state}}$ is that all models are constrained to use the same set of state clusters, as the KL-divergence requires all distributions to have the same support. Since a diversity of state cluster sets is not permitted, diverse sub-word units, context-dependencies, and HMM topologies are also not allowed. Limiting the forms of diversity allowed in the ensemble may also limit the potential combination gains.

A. State cluster diversity

Frame-level teacher-student learning can be generalised to allow for a diversity of state cluster sets, by instead minimising the KL-divergence between frame-level logical context-dependent state posteriors [12],

$$\mathcal{F}_{\text{TS}}^{\text{CD}}\left(\boldsymbol{\Theta}\right) = -\sum_{t=1}^{T}\sum_{c\in\mathcal{C}} P\left(c\middle|\mathbf{o}_t,\widehat{\boldsymbol{\Phi}}\right)\log P\left(c|\mathbf{o}_t,\boldsymbol{\Theta}\right), \quad (7)$$

where $\mathcal{C}$ represents the full set of all logical context-dependent states. However, this criterion still restricts all models to have the same HMM topology and set of sub-word units, as the set of logical context-dependent states needs to be the same across all models. The logical context-dependent state posteriors are

$$P\left(c|\mathbf{o}_t,\boldsymbol{\Theta}\right) = P\left(c\middle|s_c^{\boldsymbol{\Theta}},\mathbf{o}_t\right)P\left(s_c^{\boldsymbol{\Theta}}\middle|\mathbf{o}_t,\boldsymbol{\Theta}\right), \quad (8)$$

where $s_c^{\boldsymbol{\Theta}}$ is the student's state cluster in which logical context-dependent state c belongs. Substituting (8) into (7) gives

$$\mathcal{F}_{\text{TS}}^{\text{CD}}\left(\boldsymbol{\Theta}\right) = -\sum_{t=1}^{T}\sum_{c\in\mathcal{C}} P\left(c\middle|\mathbf{o}_t,\widehat{\boldsymbol{\Phi}}\right)\left[\log P\left(s_c^{\boldsymbol{\Theta}}\middle|\mathbf{o}_t,\boldsymbol{\Theta}\right) + \log P\left(c\middle|s_c^{\boldsymbol{\Theta}},\mathbf{o}_t\right)\right]. \quad (9)$$

If it is assumed that $P\left(c\middle|s_c^{\boldsymbol{\Theta}},\mathbf{o}_t\right)$ is independent of the student's acoustic model, then it has no influence on the student's gradient, and the criterion can be simplified to

$$\mathcal{F}_{\text{TS}}^{\text{CD}}\left(\boldsymbol{\Theta}\right) = -\sum_{t=1}^{T}\sum_{c\in\mathcal{C}} P\left(c\middle|\mathbf{o}_t,\widehat{\boldsymbol{\Phi}}\right)\log P\left(s_c^{\boldsymbol{\Theta}}\middle|\mathbf{o}_t,\boldsymbol{\Theta}\right). \quad (10)$$

This can be expressed as a sum over the student's state clusters,

$$\mathcal{F}_{\text{TS}}^{\text{CD}}\left(\boldsymbol{\Theta}\right) = -\sum_{t=1}^{T}\sum_{s^{\boldsymbol{\Theta}}\in\mathcal{T}^{\boldsymbol{\Theta}}} P\left(s^{\boldsymbol{\Theta}}\middle|\mathbf{o}_t,\widehat{\boldsymbol{\Phi}}\right)\log P\left(s^{\boldsymbol{\Theta}}\middle|\mathbf{o}_t,\boldsymbol{\Theta}\right), \quad (11)$$

where $s^{\boldsymbol{\Theta}} \in \mathcal{T}^{\boldsymbol{\Theta}}$ are all of the state clusters at the leaves of the student's decision tree.

Using a sum combination, (11) is obtained from (10) by expressing the targets as

$$P\left(s^{\boldsymbol{\Theta}}\middle|\mathbf{o}_t,\widehat{\boldsymbol{\Phi}}\right) = \sum_{m=1}^{M}\lambda_m\sum_{s^m\in\mathcal{T}^m} P\left(s^{\boldsymbol{\Theta}}\middle|s^m,\mathbf{o}_t\right)P\left(s^m|\mathbf{o}_t,\boldsymbol{\Phi}^m\right), \quad (12)$$

where $s^m \in \mathcal{T}^m$ are the state clusters at the leaves of the teachers' decision trees, and the posteriors are mapped between decision trees using

$$P\left(s^{\boldsymbol{\Theta}}\middle|s^m,\mathbf{o}_t\right) = \sum_{c:\mathcal{T}^{\boldsymbol{\Theta}}(c)=s^{\boldsymbol{\Theta}}} P\left(c|s^m,\mathbf{o}_t\right). \quad (13)$$

Here, $c:\mathcal{T}^{\boldsymbol{\Theta}}\left(c\right) = s^{\boldsymbol{\Theta}}$ are all of the logical context-dependent states that are mapped to state cluster $s^{\boldsymbol{\Theta}}$. However, the standard hybrid NN-HMM model does not capture $P\left(c|s^m,\mathbf{o}_t\right)$. Capturing this distribution is equivalent to separately modelling the observation likelihoods of each logical context-dependent state. To address this, an approximation can be made that the map is independent of the observations, $\mathbf{o}_t$,

$$P\left(s^{\boldsymbol{\Theta}}\middle|s^m,\mathbf{o}_t\right) \approx P\left(s^{\boldsymbol{\Theta}}\middle|s^m\right). \quad (14)$$

This can be computed as a discounted maximum likelihood estimate from forced alignments. However, this map is effectively a fixed linear transformation of the posteriors, which may result in a smoothing of the posteriors.

Applying this approximation to the targets in (12) gives

$$\widetilde{P}\left(s^{\boldsymbol{\Theta}}\middle|\mathbf{o}_t,\widehat{\boldsymbol{\Phi}}\right) = \sum_{m=1}^{M}\lambda_m\sum_{s^m\in\mathcal{T}^m} P\left(s^{\boldsymbol{\Theta}}\middle|s^m\right)P\left(s^m|\mathbf{o}_t,\boldsymbol{\Phi}^m\right). \quad (15)$$

These approximate targets have been shown to allow a student to learn from teachers with different sets of state clusters, over a range of different datasets [12]. Using these approximate targets, the criterion of $\mathcal{F}_{\text{TS}}^{\text{CD}}$ in (11) can now in fact be used together with not only diverse state cluster sets and context-dependencies, but also diverse HMM topologies and sets of sub-word units, as long as parallel state-level alignments exist to estimate the map, $P\left(s^{\boldsymbol{\Theta}}\middle|s^m\right)$. These latter forms of diversity shall be left for exploration in future work.

IV. Sequence teacher-student learning

The methods described in Section III train the student with frame-level criteria. As previously discussed, these impose restrictions on the forms of diversity allowed in the teacher ensemble. They also only propagate frame-level state posterior information, which does not take into account the language and alignment models, and the sequential nature of the data during training. Furthermore, previous work [13] has shown that sequence-level training often outperforms frame-level training. To overcome these limitations, teacher-student learning can be generalised to use sequence-level criteria [8].

One possible sequence-level criterion is to minimise the KL-divergence between word sequence posteriors,

$$\mathcal{F}^{\text{word}}_{\text{seq-TS}}\left(\boldsymbol{\Theta}\right) = -\sum_{\boldsymbol{\omega}} P\left(\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log P\left(\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right). \quad (16)$$

This criterion propagates sequence posterior information, which may capture the interactions between the acoustic, alignment, and languages models. A hybrid NN-HMM model is often able to produce good ASR performance [24]. This paper considers how a NN-HMM student may be trained with sequence-level teacher-student learning, where the student's hypothesis posteriors can be computed as

$$P\left(\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right) = \frac{P^{\gamma}\left(\boldsymbol{\omega}\right)\sum\limits_{\mathbf{s}^{\boldsymbol{\Theta}}\in\mathcal{G}_{\boldsymbol{\omega}}}\prod\limits_{t=1}^{T}P^{\gamma}\left(s_t^{\boldsymbol{\Theta}}\middle|s_{t-1}^{\boldsymbol{\Theta}}\right)\frac{P^{\kappa}\left(s_t^{\boldsymbol{\Theta}}|\mathbf{o}_t,\boldsymbol{\Theta}\right)}{P^{\kappa}\left(s_t^{\boldsymbol{\Theta}}\right)}}{\sum\limits_{\boldsymbol{\omega}'}P^{\gamma}\left(\boldsymbol{\omega}'\right)\sum\limits_{\mathbf{s}^{\boldsymbol{\Theta}'}\in\mathcal{G}_{\boldsymbol{\omega}'}}\prod\limits_{t=1}^{T}P^{\gamma}\left(s_t^{\boldsymbol{\Theta}'}\middle|s_{t-1}^{\boldsymbol{\Theta}'}\right)\frac{P^{\kappa}\left(s_t^{\boldsymbol{\Theta}'}|\mathbf{o}_t,\boldsymbol{\Theta}\right)}{P^{\kappa}\left(s_t^{\boldsymbol{\Theta}'}\right)}}. \quad (17)$$

Here, γ and κ are language and acoustic scaling factors that are often incorporated to tune the balance between the contributions of the language and acoustic models, and $\mathcal{G}_{\boldsymbol{\omega}}$ is the set of all state cluster sequences, $\mathbf{s}^{\boldsymbol{\Theta}}$, that correspond to the word sequence $\boldsymbol{\omega}$. Using $\mathcal{F}^{\text{word}}_{\text{seq-TS}}$, the student's gradient is shown in Appendix A to be

$$\frac{\partial\mathcal{F}^{\text{word}}_{\text{seq-TS}}}{\partial\log P\left(s_t^{\boldsymbol{\Theta}}\middle|\mathbf{o}_t,\boldsymbol{\Theta}\right)} = \kappa\left[P\left(s_t^{\boldsymbol{\Theta}}\middle|\mathbf{O},\boldsymbol{\Theta}\right) - \sum_{\boldsymbol{\omega}}P\left(s_t^{\boldsymbol{\Theta}}\middle|\boldsymbol{\omega},\mathbf{O},\boldsymbol{\Theta}\right)P\left(\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\right]. \quad (18)$$

By propagating information from the ensemble to the student in the form of word sequence posteriors, this criterion allows significant flexibility in the forms of ensemble diversity. The only requirement is that hypothesis posteriors must be obtainable from the models. This allows, for example, NN end-to-end models [25] to be used.

The first gradient term, $P\left(s_t^{\boldsymbol{\Theta}}\middle|\mathbf{O},\boldsymbol{\Theta}\right)$, can be computed in a similar fashion to the Maximum Mutual Information, $\mathcal{F}_{\text{MMI}}$, gradient denominator term, using a forward-backward operation over the student's denominator lattice [13]. The second term in the gradient requires the calculation of $P\left(s_t^{\boldsymbol{\Theta}}\middle|\boldsymbol{\omega},\mathbf{O},\boldsymbol{\Theta}\right)$ and $P\left(\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$ for every possible hypothesis, $\boldsymbol{\omega}$. These can be computed using forward-backward operations over lattices representing either the student's or teachers' hypotheses. The lattices may be pruned in a lattice-based implementation, or unpruned in a lattice-free implementation [26]. However, computing and storing these probabilities for every possible hypothesis can be computationally expensive. This cost can be limited by restricting the sum in (18) to only consider a finite n-best list of hypotheses, or by taking a Monte Carlo approximation to the gradient, similar to [27]. To the authors' knowledge, an efficient forward-backward algorithm to jointly compute $\sum_{\boldsymbol{\omega}}P\left(s_t^{\boldsymbol{\Theta}}|\boldsymbol{\omega},\mathbf{O},\boldsymbol{\Theta}\right)P\left(\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$ has not yet been proposed, and may be an interesting direction for future research.

Rather than only considering a subset of the hypotheses in the sum, an alternative solution may be to approximate the criterion, to eliminate the need to sum over sequences. One approximation is to minimise the KL-divergence between posteriors of lattice arc sequences instead of word sequences,

$$\mathcal{F}^{\text{arc}}_{\text{seq-TS}}(\boldsymbol{\Theta}) = -\sum_{\boldsymbol{\omega}}\sum_{\mathbf{a}\in\mathcal{G}_{\boldsymbol{\omega}}}P\left(\mathbf{a},\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log P(\mathbf{a},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}), \quad (19)$$

where $\mathcal{G}_{\boldsymbol{\omega}}$ is the set of all arc sequences, $\mathbf{a}$, that correspond to the word sequence $\boldsymbol{\omega}$. This criterion requires $P\left(\mathbf{a},\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$ and $P\left(\mathbf{a},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right)$ to have the same support, in that the posteriors from both the teachers and student must be defined over the same set of arc sequences. The word and arc sequence posteriors are related as

$$P\left(\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right) = \sum_{\mathbf{a}\in\mathcal{G}_{\boldsymbol{\omega}}}P\left(\mathbf{a},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right). \quad (20)$$

The difference between $\boldsymbol{\omega}$ and $\mathbf{a}$ is that arcs have defined start and end times. In addition to these times, the arcs can be marked, by incorporating in additional information about the word, sub-word unit, or state identities. However, this criterion does place constraints on the ensemble and student. Marking the arcs with sub-word units requires all models to use the same set of sub-word units. Marking the arcs with states further requires all models to use the same set of states, context-dependency, and HMM topology. The joint posterior between arc and word sequences is considered in $\mathcal{F}^{\text{arc}}_{\text{seq-TS}}$, since for certain choices of arc markings, such as state clusters or phones, the arc sequence may not uniquely determine the word sequence, because of the possibility of homophonic words. With homophonic words, the same arc sequence with different word sequences may have different language model probabilities. Since arcs have defined start and end times, the gradient of $\mathcal{F}^{\text{arc}}_{\text{seq-TS}}$ can be written as

$$\frac{\partial\mathcal{F}^{\text{arc}}_{\text{seq-TS}}}{\partial\log P\left(s_t^{\boldsymbol{\Theta}}\middle|\mathbf{o}_t,\boldsymbol{\Theta}\right)} = \kappa\left[P\left(s_t^{\boldsymbol{\Theta}}\middle|\mathbf{O},\boldsymbol{\Theta}\right) - \sum_{a_t}P\left(s_t^{\boldsymbol{\Theta}}|a_t,\mathbf{O},\boldsymbol{\Theta}\right)P\left(a_t\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\right]. \quad (21)$$

Unlike (18), the second term in (21) does not have a sum over sequences. This gradient can again be computed using two levels of forward-backward operations, one for $P\left(a_t\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$ and another for $P\left(s_t^{\boldsymbol{\Theta}}|a_t,\mathbf{O},\boldsymbol{\Theta}\right)$. These need to be computed and stored for every arc in the lattice. The number of arcs is often significantly fewer than the number of hypotheses. This gradient can again be computed within a lattice-based or lattice-free framework. However, lattice-free implementations

often only compose graphs up to the sub-word unit level to reduce the computational cost of training [26], and can therefore only be used with arcs marked with sub-word units or states.

The criterion of $\mathcal{F}^{\text{arc}}_{\text{seq-TS}}$ is in fact an upper bound to $\mathcal{F}^{\text{word}}_{\text{seq-TS}}$. To prove this, (20) can be used to express $\mathcal{F}^{\text{word}}_{\text{seq-TS}}$ in (16) as

$$\mathcal{F}^{\text{word}}_{\text{seq-TS}} = -\sum_{\boldsymbol{\omega}}\sum_{\mathbf{a}\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{a},\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log\sum_{\mathbf{a}'\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{a}',\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right). \tag{22}$$

The student's posteriors in (22) are valid probability distributions, satisfying $0 \leq P\left(\mathbf{a},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right) \leq 1$. From this, it can be seen that $\forall \mathbf{a} \in \mathcal{G}_{\boldsymbol{\omega}}$,

$$\sum_{\mathbf{a}'\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{a}',\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right) \geq P\left(\mathbf{a},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right), \tag{23}$$

and therefore

$$-\log\sum_{\mathbf{a}'\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{a}',\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right) \leq -\log P\left(\mathbf{a},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right). \tag{24}$$

Substituting (24) into (22) leads to the relationship of

$$\begin{aligned}\mathcal{F}^{\text{word}}_{\text{seq-TS}} &\leq -\sum_{\boldsymbol{\omega}}\sum_{\mathbf{a}\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{a},\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log P\left(\mathbf{a},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}\right)\\ &\leq \mathcal{F}^{\text{arc}}_{\text{seq-TS}}.\end{aligned} \tag{25}$$

Thus, minimising $\mathcal{F}^{\text{arc}}_{\text{seq-TS}}$ minimises an upper bound to $\mathcal{F}^{\text{word}}_{\text{seq-TS}}$. However, this bound may be loose, as the student's probability mass for each word sequence, $\boldsymbol{\omega}$, may be distributed among many possible arc sequences, $\mathbf{a}$, representing different arc transition times. Perhaps a tighter bound can be achieved by replacing the sum in (23) with a $\max$, and this may be an interesting direction to explore for future research.

A. State cluster sequence posteriors

The lattice arcs can be marked with a variety of acoustic units. Previous work has investigated marking the arcs with state clusters [8]. This leads to a criterion of minimising the KL-divergence between state cluster sequence posteriors,

$$\mathcal{F}^{\text{state}}_{\text{seq-TS}}(\boldsymbol{\Theta}) = -\sum_{\boldsymbol{\omega}}\sum_{\mathbf{s}\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{s},\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log P(\mathbf{s},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}). \tag{26}$$

This criterion has a gradient of

$$\frac{\partial\mathcal{F}^{\text{state}}_{\text{seq-TS}}}{\partial\log P\left(s_t|\mathbf{o}_t,\boldsymbol{\Theta}\right)} = \kappa\left[P\left(s_t|\mathbf{O},\boldsymbol{\Theta}\right) - P\left(s_t\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\right]. \tag{27}$$

The $P\left(s_t\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$ term can be computed using just a single level of forward-backward operations. In contrast to this, the equivalent term in the gradient of $\mathcal{F}^{\text{arc}}_{\text{seq-TS}}$ in (21) requires two-levels of forward-backward operations for arc markings that do not deterministically dictate the student's state cluster at each frame. This is analogous to the difference between the state cluster marked state-level MBR, $\mathcal{F}_{\text{sMBR}}$, criterion [28], [29], and the more general minimum phone or word error criteria [30]. However, using this criterion foregoes much of the freedom in the forms of ensemble diversity that are allowed by $\mathcal{F}^{\text{word}}_{\text{seq-TS}}$. In particular, all models are here restricted to use the same set of state clusters, which therefore requires identical sub-word units, HMM topologies, and context-dependencies.

B. Logical context-dependent state sequence posteriors

One possible approach to allow for more forms of diversity in the ensemble is to use the criterion of $\mathcal{F}^{\text{arc}}_{\text{seq-TS}}$ with arcs marked with words or sub-word units. However, this requires two levels of forward-backward operations for the gradient computation in (21). Furthermore, lattice-free implementations often only compose graphs up to the sub-word unit level to limit the computational cost during training [26], and thus cannot use arcs marked with words. At the frame-level, teacher-student learning can be generalised to allow for a diversity of state cluster sets, by using a criterion of minimising the KL-divergence between logical context-dependent state posteriors in $\mathcal{F}^{\text{CD}}_{\text{TS}}$ [12]. However, this requires the approximation that the map between state clusters is independent of the observation using (14), which may smooth the posteriors.

This paper proposes to generalise sequence-level teacher-student learning to allow for a diversity of state cluster sets, while still having a simple and efficient gradient computation of only a single level of forward-backward operations. Analogously to frame-level training, arcs in $\mathcal{F}^{\text{arc}}_{\text{seq-TS}}$ can be marked with logical context-dependent states, instead of state clusters. This is simpler to implement than arcs marked with words or sub-word units, and can use the lattice-free framework. The set of logical context-dependent states is common across all models, irrespective of the set of state clusters. The student can be trained by minimising the KL-divergence between logical context-dependent state sequence posteriors,

$$\mathcal{F}^{\text{CD}}_{\text{seq-TS}}(\boldsymbol{\Theta}) = -\sum_{\boldsymbol{\omega}}\sum_{\mathbf{c}\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{c},\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log P(\mathbf{c},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Theta}). \tag{28}$$

If the same set of state clusters is used across all models, it can be shown using a similar argument to (22-25) that $\mathcal{F}^{\text{word}}_{\text{seq-TS}} \leq \mathcal{F}^{\text{state}}_{\text{seq-TS}} \leq \mathcal{F}^{\text{CD}}_{\text{seq-TS}}$. Equality of $\mathcal{F}^{\text{state}}_{\text{seq-TS}} = \mathcal{F}^{\text{CD}}_{\text{seq-TS}}$ is obtained if each state cluster sequence is uniquely described by a single logical context-dependent state sequence. This can be achieved, for example, by having a different root in the decision trees for each different centre phone.

The gradient of $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$ can be expressed as

$$\frac{\partial\mathcal{F}^{\text{CD}}_{\text{seq-TS}}}{\partial\log P\left(s_t^{\boldsymbol{\Theta}}|\mathbf{o}_t,\boldsymbol{\Theta}\right)} = \kappa\left[P\left(s_t^{\boldsymbol{\Theta}}\middle|\mathbf{O},\boldsymbol{\Theta}\right) - \sum_{\widehat{s}_t\in\mathcal{G}_{s_t^{\boldsymbol{\Theta}}}} P\left(\widehat{s}_t\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\right], \tag{29}$$

where the "intersect states", $\widehat{s}_t$, define the set of state clusters formed by the Cartesian product of all decision trees of the teachers and student [31]. Here, $\mathcal{G}_{s_t^{\boldsymbol{\Theta}}}$ represents the set of intersect states whose constituent logical context-dependent states are tied to the student's state cluster $s_t^{\boldsymbol{\Theta}}$. The $P\left(\widehat{s}_t\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$ term can be computed efficiently using a single level of forward-backward operations. As opposed to this, the equivalent term in (21) requires two levels of forward-backward operations to compute. It is thus simpler to implement and possibly more efficient to mark the arcs with logical context-dependent states, rather than words or sub-word units, when allowing for a diversity of state cluster sets. The arcs in fact need not be marked with logical context-dependent states. It is more efficient to mark them with intersect states. When using

pruned lattices in a lattice-based implementation, it again has to be ensured that the same intersect state sequences exist in the lattices used for all of the models.

In the frame-level targets of (12), the map, $P\left(s^{\boldsymbol{\Theta}}\middle|s^m,\mathbf{o}_t\right)$, is approximated using (14). This approximation may smooth the target posteriors. At the sequence level, marking the arcs with logical context-dependent states allows the targets of $P\left(\mathbf{c},\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$ in the $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$ criterion to be computed exactly, without the need for any approximations. The surrounding context of a phone is known when given the state cluster sequence. This avoids any performance degradation of the student that may result from any approximations.

The criterion proposed in this paper generalises teacher-student learning to allow for sequence-level training when the ensemble has a diversity of state cluster sets, while preserving the simplicity and efficiency of only requiring a single level of forward-backward operations when computing the gradient. However, this method still constrains all models to use the same set of sub-word units and HMM topology.

C. Sequence posterior targets

In these sequence-level teacher-student criteria, there are several possible methods of combining the posteriors from the teachers in the ensemble to obtain the targets. For $\mathcal{F}^{\text{state}}_{\text{seq-TS}}$, a sum combination of the sequence posteriors can be used [8],

$$P\left(\mathbf{s},\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)=\sum_{m=1}^{M}\lambda_m P\left(\mathbf{s},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Phi}^m\right). \tag{30}$$

With this form of target combination, the $P\left(s_t\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$ term in the gradient of (27) can be computed as a sum over the contributions of the denominator lattices from each teacher,

$$P\left(s_t\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)=\sum_{m=1}^{M}\lambda_m P\left(s_t|\mathbf{O},\boldsymbol{\Phi}^m\right). \tag{31}$$

It is also possible to take a product of the sequence posteriors [14], which for hybrid NN-HMM models, is equivalent to taking a frame-level product of the teachers' acoustic scores,

$$\begin{aligned}P\left(\mathbf{s},\boldsymbol{\omega}\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)&=\frac{1}{Z}\prod_{m=1}^{M}P^{\lambda_m}\left(\mathbf{s},\boldsymbol{\omega}|\mathbf{O},\boldsymbol{\Phi}^m\right) \qquad (32)\\ &=\frac{1}{Z}P\left(\boldsymbol{\omega}\right)P\left(\mathbf{s}\right)\prod_{t=1}^{T}\prod_{m=1}^{M}p^{\kappa\lambda_m}\left(\mathbf{o}_t|s_t,\boldsymbol{\Phi}^m\right),\end{aligned}$$

assuming that $\sum_m \lambda_m = 1$, and both the language and alignment models are the same across all teachers. Here, Z ensures that the combination results in a normalised distribution. With this form of target combination, only a single denominator lattice needs to be processed for all of the teachers to compute $P\left(s_t\middle|\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$ in the gradient. This denominator lattice is generated with a frame-level combination of the teachers' acoustic scores. During training, it is therefore less computationally expensive to use a product combination to obtain the targets, rather than a hypothesis-level sum combination.

This paper compares a range of frame and sequence-level criteria for teacher-student learning that operate on different levels of acoustic units. A summary of the notation used for these approaches is shown in Table I.

TABLE I
SUMMARY OF FRAME AND SEQUENCE-LEVEL TEACHER-STUDENT LEARNING CRITERIA, COMPUTED OVER DIFFERENT ACOUSTIC UNITS

	Acoustic unit			
Criterion	word	arc	state cluster	logical state
frame-level	-	-	$\mathcal{F}^{\text{state}}_{\text{TS}}$	$\mathcal{F}^{\text{CD}}_{\text{TS}}$
sequence-level	$\mathcal{F}^{\text{word}}_{\text{seq-TS}}$	$\mathcal{F}^{\text{arc}}_{\text{seq-TS}}$	$\mathcal{F}^{\text{state}}_{\text{seq-TS}}$	$\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$

V. EXPERIMENTS

The experiments are divided into four sections. Section V-A first assesses the proposed sequence-level teacher-student criterion of $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$, which allows for different sets of state clusters between the teachers, comparing its performance to frame-level training and sequence-level teacher-student learning with the same set of state clusters. Next, Section V-B considers a lattice-free implementation of teacher-student learning. Section V-C then investigates the possibility of incorporating additional model parameter diversity from within each training run. Finally, Section V-D assesses the effectiveness of different schemes for compressing such an ensemble.

The experiments were performed using the Kaldi speech recognition toolkit [32]. Two datasets were used. The AMI meeting transcription task [33] 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 [34] comprises audio recordings from television programs of a variety of genres. Lightly supervised decoding and selection [35] 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 [36] that was trained on the MGB-3 data.

For both datasets, 40-dimensional Mel-scale filterbank features were extracted. Experiments used models trained with both lattice-based and lattice-free implementations of sequence training. For the lattice-free models, biphone decision trees were built with approximately 2000 and 3600 leaves for AMH-IHM and MGB-3 respectively. A lattice-free implementation of the $\mathcal{F}_{\text{MMI}}$ criterion was used to train models with interleaved Time Delay NN (TDNN) and Long Short-Term Memory (LSTM) layers [11], referred to as TDNN-LSTM. The TDNN layers had 600 nodes with rectified linear unit activations, and the LSTM layers had 512 cells with 128 recurrent and non-recurrent projections. The topology can be described as $\{-2,-1,0,1,2\}$ $\{-1,0,1\}$ L $\{-3,0,3\}$ $\{-3,0,3\}$ L $\{-3,0,3\}$ $\{-3,0,3\}$ L, where L represent a LSTM layer. These models were trained using an exponential learning rate schedule, where the learning rates and number of epochs were selected using hyper-parameter sweeps. Models trained in a lattice-based framework were also used for the experiments in Section V-A for AMI-IHM. For these, alignments were first obtained from a Gaussian Mixture Model (GMM)-HMM following the Kaldi *s5* recipe, and triphone decision trees were built with 4000 leaves. These were used to train feed-forward sigmoid Deep NNs (DNN) with 6 layers of 2048

nodes. First and second temporal derivatives were appended to the input features of these DNNs, with an 11 frame context. The DNNs were first initialised with layer-wise supervised pretraining, followed by Cross-Entropy, $\mathcal{F}_{\text{CE}}$, training, then lattice-based sequence training using the $\mathcal{F}_{\text{sMBR}}$ criterion. Evaluation was done using MBR decoding, and hypothesis-level MBR combination decoding, defined in (2), was used for ensemble combination [6]. 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. Equal interpolation weights, $\lambda_m = \frac{1}{M}$ and $\eta_m = \frac{1}{M}$, were used for MBR combination decoding, teacher-student learning, and parameter-level combination. Preliminary experiments suggested that training these weights to optimise the $\mathcal{F}_{\text{MMI}}$ criterion in parameter-level combination did not result in any significant gains over using equal weights.

A. Frame and sequence-level teacher-student learning

TABLE II
LATTICE-BASED SMBR DNN ENSEMBLES, IN AMI-IHM

Diversity	Single WER (%) mean	Single WER (%) std dev	Combined WER (%)	cross-WER (%)
parameter	25.7	0.10	24.9	11.8
state cluster	26.0	0.13	24.5	15.2

The sequence-level $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$ criterion proposed in this paper to allows a student to learn from an ensemble with a diversity of state cluster sets. The first experiment compares the performances of ensembles generated with either this form of diversity or a diversity of model parameters. An ensemble with model parameter diversity was generated by training 4 DNNs with the same greedy decision tree, separately toward a lattice-based implementation of the $\mathcal{F}_{\text{sMBR}}$ criterion, each starting from a different random parameter initialisation. An ensemble with a diversity of state cluster sets was generated by training 4 DNNs separately toward the $\mathcal{F}_{\text{sMBR}}$ criterion, each with a different decision tree. Multiple decision trees were obtained using the random forest method, by sampling splits from the 5-best at each iteration. The performances of these ensembles are shown in Table II. The results show that diversity and performance gains can be obtained in both ensembles. The diversity between the models was estimated using the cross-WER [37]. This measures the word-level minimum edit distance between the 1-best hypotheses of each model within an ensemble, averaged across all pairs of models. A larger cross-WER indicates greater diversity between the model predictions. The ensemble with a diversity of state cluster sets exhibits a wider diversity and has a better combined performance than the ensemble with model parameter diversity, with a null hypothesis probability less than 0.001, using the matched pairs sentence segment word error test [38]. The single model performance of this ensemble may be worse because the random forest decision trees may be less optimal.

TABLE III
FRAME AND LATTICE-BASED SEQUENCE-LEVEL TEACHER-STUDENT LEARNING, IN AMI-IHM

Ensemble	Training	Student WER (%)
-	$\mathcal{F}_{\text{CE}} + \mathcal{F}_{\text{sMBR}}$	25.7
parameter	$\mathcal{F}^{\text{state}}_{\text{TS}}$	25.1
	$\mathcal{F}^{\text{state}}_{\text{seq-TS}}$	24.7
state cluster	$\mathcal{F}^{\text{CD}}_{\text{TS}}$	25.5
	$\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$	24.6

However, these ensembles require multiple decoding runs and are therefore computationally expensive to use for recognition. To reduce this cost, students were trained toward the ensemble with model parameter diversity using the frame-level $\mathcal{F}^{\text{state}}_{\text{TS}}$ and sequence-level $\mathcal{F}^{\text{state}}_{\text{seq-TS}}$ criteria, and toward the ensemble with state cluster diversity using the frame-level $\mathcal{F}^{\text{CD}}_{\text{TS}}$ [12] and proposed sequence-level $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$ criteria. These teacher-student methods are compared in Table III. The student models used the same DNN topology as each teacher in the ensemble, but with a decision tree that had been trained using greedy splits. The sequence-level students here were trained using lattice-based implementations of the gradient computations, and used the frame-level students as parameter initialisations. To ensure the same support for the sequence posterior distributions of the teachers and student, the denominator lattices for the teachers were obtained by rescoring the student's denominator lattice paths separately with each of the teachers' acoustic scores. Obtaining the denominator lattice paths from the initial frame-level students allowed for a reasonable match of these paths with all of the teachers' and student's acoustic models. The arcs in the denominator lattices for $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$ were marked with intersect states, to allow for the different sets of state clusters. The results show that all criteria are able to bring the student performance closer to that of the combined ensemble in Table II, than when using $\mathcal{F}_{\text{sMBR}}$ training. The sequence-level students are able to come closer to the ensemble performances than the frame-level students. As such, the sequence posterior information propagated over by the proposed $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$ criterion may allow the student to learn more effectively from teachers with different sets of state clusters, than propagating only frame posterior information.

Although the ensemble with different state cluster sets has a better combined performance, in Table II, its frame-level student does not perform as well as that of the ensemble with model parameter diversity. Sequence-level teacher-student learning reduces the difference between the student performances. However, the sequence-level student of the ensemble with state cluster diversity is not significantly better than that of the ensemble with model parameter diversity, with a null hypothesis probability of 0.144. Work in [12] proposes to improve the student performance by using a larger decision tree, and this method is demonstrated for the lattice-free experiments in Section V-B.

As is shown in [8], it is also possible to use standard sequence training to further improve upon the frame-level student performance. Further $\mathcal{F}_{\text{sMBR}}$ training of the $\mathcal{F}^{\text{CD}}_{\text{TS}}$ student yields a WER of 24.6%, which is similar to the sequence-level

student performance. This also outperforms an $\mathcal{F}_{\text{sMBR}}$ model trained from $\mathcal{F}_{\text{CE}}$ initialisation, suggesting that the frame-level student may be a better initialisation for sequence training.

B. Lattice-free sequence-level teacher-student learning

TABLE IV
LATTICE-FREE MMI TDNN-LSTM ENSEMBLES WITH DIFFERENT SETS OF STATE CLUSTERS

Dataset	Single WER (%) mean	Single WER (%) std dev	Combined WER (%)	cross-WER (%)
AMI-IHM	25.6	0.06	22.3	21.3
MGB-3	23.5	0.18	20.6	18.2

The experiments in Section V-A show that an ensemble with different sets of state clusters can exhibit significant combination gains, and that a lattice-based implementation of the proposed sequence-level method can allow a student to learn from such an ensemble. This section examines a lattice-free implementation of the proposed sequence-level teacher-student learning criterion. The lattice-free framework [26] allows for sequence training without needing to prune lattices. This removes the need of an initial frame-level model to provide acoustic scores for lattice pruning, and therefore allows sequence training to begin from a random parameter initialisation. Ensembles of 4 models were trained toward a lattice-free implementation of the $\mathcal{F}_{\text{MMI}}$ criterion, again with a different set of state clusters for each model. The TDNN-LSTM topology was used for the lattice-free models, as this was found to outperform the feed-forward DNN topology. Table IV shows the performances of these lattice-free ensembles. Comparing this with Table II, the results suggest that the lattice-free ensemble is able to exhibit a greater diversity than the lattice-based ensemble. As a reference, a lattice-based $\mathcal{F}_{\text{sMBR}}$ ensemble of TDNN-LSTMs has a combined WER of 22.2% and a cross-WER of 18.6% in AMI-IHM. Training the ensemble using a lattice-based implementation of the $\mathcal{F}_{\text{MMI}}$ criterion, instead of $\mathcal{F}_{\text{sMBR}}$, does not yield any significant increase in the diversity. As such, the increase in the lattice-free ensemble diversity in Table IV does not appear to be solely due to the different model topology or training criterion used. In lattice-based training, the $\mathcal{F}_{\text{CE}}$ initialisation may strongly bias the models toward the forced alignments. On the other hand, lattice-free training can begin from a random parameter initialisation. Although alignment information is still used to split the utterances into shorter segments for training and to compute the numerator lattice, this information is weakened by allowing for state transitions within a window around the alignments. This may reduce the bias that the lattice-free models have toward the forced alignments, and allow them to develop more diverse behaviours. However, the lattice-free ensemble with a combined WER of 22.3% does not outperform a lattice-based TDNN-LSTM ensemble with a combined WER of 22.2%, in AMI-IHM. Despite this, it is still advantageous to use lattice-free training, as it is computationally cheaper during both training and recognition. This is because lattice-free training can begin from random parameter initialisation and a lower frame rate is used.

TABLE V
LATTICE-FREE SEQUENCE-LEVEL TEACHER-STUDENT LEARNING ACROSS DIFFERENT SETS OF STATE CLUSTERS

Dataset	Training	WER (%)
AMI-IHM	$\mathcal{F}_{\text{MMI}}$	25.3
	$\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$	23.2
MGB-3	$\mathcal{F}_{\text{MMI}}$	23.6
	$\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$	21.8

Using a lattice-free implementation of $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$, students were trained toward these lattice-free ensembles. The students used the same TDNN-LSTM topology as each of the teachers in the ensembles, but with decision trees that were trained with greedy splits. These students were trained, beginning from random parameter initialisations. The results in Table V show that the students are able to come closer to the ensemble performances than when using lattice-free $\mathcal{F}_{\text{MMI}}$ training.

The targets used to train the students have thus far been combined using a sum combination of (30). It is also possible to use a product combination of (32) [14]. A student trained using a product combination has a WER of 23.5% in AMI-IHM. The student using a sum combination with a WER of 23.2% in Table V appears to perform slightly better. However, this may not be statistically significant, with a null hypothesis probability of 0.010. It is cheaper to compute the gradient using a product combination, as only a single denominator lattice needs to be processed for the whole ensemble. The remaining experiments use a sum combination.

Although the students' performances approach those of the ensembles, they are still significantly different. A similar degradation of the student is observed when performing frame-level teacher-student learning using the $\mathcal{F}^{\text{CD}}_{\text{TS}}$ criterion in [12]. In [12], it is proposed that two possible sources of performance degradation are the approximation used to map the frame-level posterior targets across different sets of state clusters, and the limited phonetic resolution of the student. The student can produce separate acoustic scores for each state cluster within its decision tree, while the combined teachers can effectively produce separate acoustic scores for each intersect state within the Cartesian product of all of their decision trees. When training the student using $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$, no approximations are needed to obtain the targets. Therefore, the performance degradation of the students here may be primarily due to their limited phonetic resolutions, as they used standard-sized decision trees. In AMI-IHM, a $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$ student that used the intersect states has a WER of 22.5%. These intersect states were obtained from a Cartesian product of the 4 decision trees of the teachers, and therefore has the same phonetic resolution as the ensemble. This student comes closer to the combined ensemble performance of 22.3%. However, the intersect student with 11581 state clusters, has 12.1×10^6 parameters, which is many more than the 9.6×10^6 parameters in a student using a standard-sized decision tree with 2000 state clusters. This may increase the computational cost when performing recognition. It may be possible to use the multi-task topology proposed in [37] to retain the ensemble's phonetic resolution, while reducing the number of model parameters.

C. Ensembles with smoothed models

The ensembles used in the previous experiments have a diversity of state cluster sets. It may be possible to further improve the ensemble performance by incorporating additional forms of diversity, such as a diversity of model parameter sets. One possible method for obtaining model parameter diversity is to construct an ensemble from the intermediate model iterations within a single run of training [15]. This does not require any additional computation during training, and can be seamlessly integrated with an ensemble that already has a diversity of state cluster sets. The next experiment assesses ensembles that were generated using these diversity methods. Ensembles were generated from within a single run of lattice-free $\mathcal{F}_{\text{MMI}}$ training, using a greedy decision tree. To limit the computational cost, every 3rd and 12th model from the last training epoch was used in AMI-IHM and MGB-3 respectively, leading to ensembles of 20 and 22 models. These will be referred to as Single Run (SR) ensembles. The lattice-free ensembles in Table IV that used the last model iterations of multiple training runs with different sets of state clusters will be referred to as Random Forest (RF) ensembles. One method of using model parameter and state cluster diversity together, referred to as $\text{RF}_{\text{smooth}}$, is to first perform parameter-level combination over each training run, then combine the smoothed models over the multiple training runs. This is computationally cheaper than directly combining the intermediate models across the multiple training runs. Such parameter-level combination within a training run has been shown to yield significant performance gains over using just the last model iteration [11].

TABLE VI
RANDOM FOREST AND SINGLE RUN ENSEMBLE METHODS

Dataset	Ensemble Method	Single WER (%) mean	Single WER (%) std dev	Combined WER (%)	cross-WER (%)
AMI-IHM	SR	25.6	0.34	23.5	15.6
	RF	25.6	0.06	22.3	21.3
	$\text{RF}_{\text{smooth}}$	24.2	0.06	21.6	18.2
MGB-3	SR	24.0	0.34	20.8	16.9
	RF	23.5	0.18	20.6	18.2
	$\text{RF}_{\text{smooth}}$	21.3	0.08	19.7	12.8

Table VI compares these ensembles. Again, hypothesis-level MBR combination decoding was used. The results show that significant diversity and combination gains can be obtained by generating an ensemble from the intermediate model iterations within a single run of training, in the SR ensemble, though less than when using a diversity of state cluster sets. The last model iterations in the SR training runs have WERs of 25.3 and 23.6% for AMI-IHM and MGB-3 respectively, from Table V. Performing parameter-level combination within each training run leads to improved single model performances in the $\text{RF}_{\text{smooth}}$ ensembles, compared to the RF ensembles. This in turn yields a better combined performance, with null hypothesis probabilities less than 0.001 for both datasets. However, this also leads to a reduction in the diversity between the smoothed models across the separate training runs.

The hypothesis-level combined performances of the SR ensembles, of 23.5 and 20.8% for AMI-IHM and MGB-3 respectively, are better than the mean performances of the separate smoothed models in the $\text{RF}_{\text{smooth}}$ ensembles of 24.2 and 21.3%. This may be partly attributed to the random forest decision trees in $\text{RF}_{\text{smooth}}$, compared to the greedy decision trees in SR. These results also suggest that it may be interesting to compare different methods of combining the intermediate models of a single training run. The SR ensemble in AMI-IHM has WERs of 23.8% for parameter-level combination and 22.6% for a student using sequence-level teacher-student learning with $\mathcal{F}^{\text{state}}_{\text{seq-TS}}$. The WERs for MGB-3 are 21.3% for parameter-level combination and 21.3% for a student. These results, compared with Table VI, suggest that hypothesis-level combination of the SR ensemble significantly outperforms parameter-level combination, with null hypothesis probabilities less than 0.001 for both datasets. In AMI-IHM, the student gives the best performance of all of the combination methods, but this is not replicated in MGB-3.

D. Multi-stage compression

The previous experiment shows how additional parameter diversity can be incorporated into an ensemble together with a diversity of state cluster sets, simply by using the intermediate models within each training run. However, the resulting ensemble can be large, leading to a high computational cost when performing recognition. The $\text{RF}_{\text{smooth}}$ ensemble in Table VI is an example of using a 2-stage combination scheme to reduce this cost, by first performing parameter-level combination over each training run, followed by hypothesis-level combination over the smoothed models from the separate training runs. This section compares this with various other schemes for compressing the ensemble into a single student model.

A naive compression scheme is to train a student directly toward all of the intermediate models across all training runs. It is interesting to compare this with a 2-stage approach, that first compresses within each training run, then compresses across the multiple training runs. In a 2-stage scheme, multiple models from a single training run can be first compressed into a single model, by using either parameter-level combination or teacher-student learning. The results in Section V-C suggest that in both cases, the resulting single model outperforms the models that make up the ensemble. Teacher-student learning can then be used to compress these single models across the multiple training runs into a final student model.

An ensemble was generated from 4 training runs, each with a different decision tree. Intermediate model iterations within the last epoch of each training run were incorporated into the ensemble. Table VII compares training a student toward all of the intermediate model iterations across all training runs, against a 2-stage compression scheme of first combining within each training run using parameter-level combination or sequence-level teacher-student learning with $\mathcal{F}^{\text{state}}_{\text{seq-TS}}$, then training a final student toward the smoothed or student models across the separate training runs using $\mathcal{F}^{\text{CD}}_{\text{seq-TS}}$. The left three columns report the mean WER, standard deviation, and cross-WER diversity across the different training runs, after the first stage of compression within each training run. The right two columns report the performance after the second stage of compression across the multiple training runs.

TABLE VII
MULTI-STAGE COMPRESSION METHODS

Stage 1 combination	Stage 1 WER (%) mean	std dev	Stage 1 cross-WER (%)	Stage 2 WER (%) hypothesis	student
AMI-IHM					
–	25.9	0.34	20.6	21.8	22.9
parameter	24.2	0.06	18.2	21.6	22.7
student	22.8	0.10	14.1	21.3	23.1
MGB-3					
–	24.0	0.41	18.6	20.0	21.5
parameter	21.3	0.08	12.8	19.7	21.2
student	21.3	0.06	10.0	20.1	21.9

The results suggest that the best final student can be obtained by first performing parameter-level combination within each training run, then performing teacher-student learning toward these smoothed models across the multiple training runs. However, these students may not be significantly better than students trained directly toward all of the intermediate model iterations across the multiple training runs, without a stage 1 compression, with null hypothesis probabilities of 0.204 and 0.016 for AMI-IHM and MGB-3 respectively. Despite this, there is a consistent improvement across both datasets. The 2-stage compression is also less computationally expensive, as this trains a student toward fewer teachers. Parameter-level combination has negligible computational cost, as the interpolation weights were not trained.

Comparing the stage 1 compression methods, performing teacher-student learning within each training run yields a better performance than parameter-level combination in AHI-IHM, but this trend is not replicated in MGB-3. A stage 1 compression using teacher-student learning yields less diversity between the students of the different training runs, than between the smoothed models with parameter-level combination. This lack of diversity may be a reason why a final student trained toward the stage 1 students does not perform as well as one trained toward the stage 1 parameter-level combinations.

Although the stage 2 hypothesis-level combinations are able to gain from both forms of diversity, the best stage 2 students are not able to significantly outperform the stage 1 students. The stage 1 students used the same sets of state clusters as their teachers, while the stage 2 students were trained toward teachers with different sets of state clusters. The stage 2 students may again be limited by their phonetic resolutions, as these students used the same-sized decision trees as each of the teachers. Using the intersect states for the stage 2 students after a stage 1 parameter-level combination yields WERs of 22.1 and 20.8% for AMI-IHM and MGB-3 respectively. This suggests that the students can benefit from the diversity provided by having different sets of state clusters and different sets of model parameters from intermediate training iterations.

VI. CONCLUSION

This paper has presented a generalisation of the teacher-student learning framework, that allows for different sets of state clusters to be used between the teacher and student models, while using a sequence-level criterion. The experimental results have shown that using the proposed method, the student can be trained to effectively emulate the combined performance of an ensemble with a diversity of state cluster sets. Also, sequence-level teacher-student learning can yield a better student performance than using frame-level criteria. Additional model parameter diversity can be incorporated into an ensemble by using intermediate model iterations within each training run. Such an ensemble can be effectively compressed using a multi-stage scheme.

The proposed sequence-level teacher-student learning criterion allows the ensemble to capture a diversity of state cluster sets. This has been shown to be an upper bound to an even more general criterion of minimising the KL-divergence between word sequence posteriors, which allows for more forms of diversity within the ensemble. Investigating efficient methods to compute the gradient of this criterion may be an interesting direction for future research.

APPENDIX A
SEQUENCE TEACHER-STUDENT CRITERION GRADIENT

A simple way to derive the gradient of the $\mathcal{F}^{\text{word}}_{\text{seq-TS}}$ criterion in (16) is to relate it to the standard $\mathcal{F}_{\text{MMI}}$ criterion as

$$\mathcal{F}^{\text{word}}_{\text{seq-TS}}\left(\boldsymbol{\Theta}\right) = \sum_{\boldsymbol{\omega}} P\left(\boldsymbol{\omega}\middle|\mathbf{O}, \widehat{\boldsymbol{\Phi}}\right) \mathcal{F}_{\text{MMI}}\left(\boldsymbol{\Theta}, \boldsymbol{\omega}\right), \tag{33}$$

where

$$\mathcal{F}_{\text{MMI}}\left(\boldsymbol{\Theta}, \boldsymbol{\omega}\right) = -\log P\left(\boldsymbol{\omega}|\mathbf{O}, \boldsymbol{\Theta}\right). \tag{34}$$

The standard $\mathcal{F}_{\text{MMI}}$ gradient is [13]

$$\frac{\partial \mathcal{F}_{\text{MMI}}\left(\boldsymbol{\Theta}, \boldsymbol{\omega}\right)}{\partial \log p\left(s_t^{\boldsymbol{\Theta}}|\mathbf{o}_t, \boldsymbol{\Theta}\right)} = \kappa\left[P\left(s_t^{\boldsymbol{\Theta}}|\mathbf{O}, \boldsymbol{\Theta}\right) - P\left(s_t^{\boldsymbol{\Theta}}|\boldsymbol{\omega}, \mathbf{O}, \boldsymbol{\Theta}\right)\right]. \tag{35}$$

Using the chain rule, the gradient of $\mathcal{F}^{\text{word}}_{\text{seq-TS}}$ is

$$\begin{aligned}\frac{\partial \mathcal{F}^{\text{word}}_{\text{seq-TS}}}{\partial \log p\left(s_t^{\boldsymbol{\Theta}}|\mathbf{o}_t, \boldsymbol{\Theta}\right)} =& \sum_{\boldsymbol{\omega}} \frac{\partial \mathcal{F}^{\text{word}}_{\text{seq-TS}}}{\partial \mathcal{F}_{\text{MMI}}\left(\boldsymbol{\Theta}, \boldsymbol{\omega}\right)} \frac{\partial \mathcal{F}_{\text{MMI}}\left(\boldsymbol{\Theta}, \boldsymbol{\omega}\right)}{\partial \log p\left(s_t^{\boldsymbol{\Theta}}|\mathbf{o}_t, \boldsymbol{\Theta}\right)} \\ =& \kappa\left[P\left(s_t^{\boldsymbol{\Theta}}|\mathbf{O}, \boldsymbol{\Theta}\right)\right. \\ & \left. - \sum_{\boldsymbol{\omega}} P\left(s_t^{\boldsymbol{\Theta}}|\boldsymbol{\omega}, \mathbf{O}, \boldsymbol{\Theta}\right) P\left(\boldsymbol{\omega}\middle|\mathbf{O}, \widehat{\boldsymbol{\Phi}}\right)\right].\end{aligned} \tag{36}$$

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] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, Montréal, Canada, Dec 2014.

[5] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," in *ASRU*, Santa Barbara, USA, Dec 1997, pp. 347–354.

[6] H. Xu *et al.*, "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.

[7] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, Oct 1990.

[8] 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.
[9] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *Interspeech*, Singapore, Sep 2014, pp. 1915–1919.
[10] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *ICASSP*, Philadelphia, USA, Mar 2005, pp. 197–200.
[11] Y. Wang *et al.*, "Phonetic and graphemic systems for multi-genre broadcast transcriptions," in *ICASSP*, Calgary, Canada, Apr 2018.
[12] 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.
[13] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *ICASSP*, Taipei, Apr 2009, pp. 3761–3764.
[14] 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.
[15] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for WMT 16," in *Machine Translation*, Berlin, Germany, Aug 2016, pp. 371–376.
[16] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *HLT*, Plainsboro, USA, Mar 1994, pp. 307–312.
[17] L. Hyafil and R. L. Rivest, "Constructing optimal binary decision trees is NP-complete," *Information Processing Letters*, vol. 5, no. 1, pp. 15–17, May 1976.
[18] T. Zhao, Y. Zhao, and X. Chen, "Building an ensemble of CD-DNN-HMM acoustic model using random forests of phonetic decision trees," in *ISCSLP*, Singapore, Sep 2014, pp. 98–102.
[19] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, Aug 2000.
[20] C. Breslin and M. J. F. Gales, "Complementary system generation using directed decision trees," in *ICASSP*, Honolulu, USA, Apr 2007, pp. 337–340.
[21] K. Kirchhoff and J. A. Bilmes, "Combination and joint training of acoustic classifiers for speech recognition," in *Automatic Speech Recognition: challenges for the new millenium*, Paris, France, Sep 2000.
[22] J. Utans, "Weight averaging for neural networks and local resampling schemes," in *AAAI*, Portland, USA, Aug 1996.
[23] H. Chen, S. Lundberg, and S.-I. Lee, "Checkpoint ensembles: ensemble methods from a single training process," Oct 2017, arXiv preprint arXiv:1710.03282.
[24] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML*, Beijing, China, Jun 2014, pp. 1764–1772.
[25] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML Representation Learning Workshop*, Edinburgh, UK, Jul 2012.
[26] D. Povey *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, San Francisco, USA, Sep 2016, pp. 2751–2755.
[27] M. Shannon, "Optimizing expected word error rate via sampling for speech recognition," in *Interspeech*, Stockholm, Sweden, Aug 2017, pp. 3537–3541.
[28] M. Gibson and T. Hain, "Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition," in *Interspeech*, Pittsburgh, USA, Sep 2006, pp. 2406–2409.
[29] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to MPE for large scale discriminative training," in *ICASSP*, Honolulu, USA, Apr 2007, pp. 321–324.
[30] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *ICASSP*, Orlando, USA, May 2002, pp. 105–108.
[31] J. Xue and Y. Zhao, "Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 3, pp. 519–528, Mar 2008.
[32] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, Hawaii, USA, Dec 2011.
[33] J. Carletta *et al.*, "The AMI meeting corpus: a pre-announcement," in *MLMI*, Edinburgh, UK, July 2005, pp. 28–39.
[34] P. Bell, "MGB challenge," May 2017, http://www.mgb-challenge.org/english.html.
[35] P. Lanchantin *et al.*, "Selection of multi-genre broadcast data for the training of automatic speech recognition systems," in *Interspeech*, San Francisco, USA, Sep 2016, pp. 3057–3061.
[36] L. Wang *et al.*, "Improved DNN-based segmentation for multi-genre broadcast audio," in *ICASSP*, Shanghai, China, May 2016, pp. 5700–5704.
[37] 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.
[38] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *ICASSP*, Glasgow, UK, May 1989, pp. 532–535.

Jeremy Wong received a BE in Electrical Engineering and a BSc in Physics from the University of New South Wales in 2011. He completed a PhD at the University of Cambridge in 2019, under the supervision of Professor Mark Gales, and was funded by the National Science Scholarship in Singapore. He was a research engineer in the Institute for Infocomm Research in Singapore, working on computer vision in 2011 and automatic speech recognition in 2014. He is currently at Microsoft, researching automatic speech recognition and speaker diarisation.

Mark Gales studied for the B.A. in Electrical and Information Sciences at the University of Cambridge from 1985-88. Following graduation he worked as a consultant at Roke Manor Research Ltd. In 1991 he took up a position as a Research Associate in the Speech Vision and Robotics group in the Engineering Department at Cambridge University. In 1995 he completed his doctoral thesis: Model-Based Techniques for Robust Speech Recognition supervised by Professor Steve Young. From 1995-1997 he was a Research Fellow at Emmanuel College Cambridge. He was then a Research Staff Member in the Speech group at the IBM T.J.Watson Research Center until 1999 when he returned to Cambridge University Engineering Department as a University Lecturer. He was appointed Reader in Information Engineering in 2004. He is currently a Professor of Information Engineering and a College Lecturer and Official Fellow of Emmanuel College. Mark Gales is a Fellow of the IEEE and a Senior Area Editor of IEEE/ACM Transactions on Audio Speech and Language Processing for speech recognition and synthesis. He was an associate editor for IEEE Signal Processing Letters from 2008-2011, IEEE Transactions on Audio Speech and Language Processing from 2009-2013 and a member of the Speech and Language Processing Technical Committee (2015-2017 and 2001-2004). He is currently on the Editorial Board of Computer Speech and Language.

Mark Gales has been awarded a number of paper awards, including a 1997 IEEE Young Author Paper Award for his paper on Parallel Model Combination and a 2002 IEEE Paper Award for his paper on Semi-Tied Covariance Matrices.

Yu Wang is a Research Associate in the speech group of Machine Intelligence Laboratory in the Engineering Department, University of Cambridge. He received the Bachelors degree from Huazhong University of Science and Technology, Wuhan, China, in 2009, the M.Sc. degree in communications and signal processing and the Ph.D. degree in signal processing, both from Imperial College, London, U.K. in 2010 and 2015, respectively. He was a Research intern at Nuance Communications from April to August 2014. Since August 2015, he has been working as a Research Associate at the Cambridge University Engineering Department on the Automated Language Teaching and Assessment (ALTA) project. His current research interests include robust speech recognition, speech and audio signal processing and automatic spoken language assessment.