HMM VS. CTC FOR AUTOMATIC SPEECH RECOGNITION: COMPARISON BASED ON FULL-SUM TRAINING FROM SCRATCH

Tina Raissi [⋆1], *Wei Zhou* [⋆1,2], *Simon Berger* [1,2], *Ralf Schlüter* [1,2], *Hermann Ney* [1,2]

[1]Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Germany
[2]AppTek GmbH, Aachen, Germany
{*raissi,zhou,sberger,schlueter,ney*}@*cs.rwth-aachen.de*

ABSTRACT

In this work, we compare from-scratch sequence-level cross-entropy (full-sum) training of Hidden Markov Model (HMM) and Connectionist Temporal Classification (CTC) topologies for automatic speech recognition (ASR). Besides accuracy, we further analyze their capability for generating high-quality time alignment between the speech signal and the transcription, which can be crucial for many subsequent applications. Moreover, we propose several methods to improve convergence of from-scratch full-sum training by addressing the alignment modeling issue. Systematic comparison is conducted on both Switchboard and LibriSpeech corpora across CTC, posterior HMM with and w/o transition probabilities, and standard hybrid HMM. We also provide a detailed analysis of both Viterbi forced-alignment and Baum-Welch full-sum occupation probabilities.

Index Terms— ASR, HMM, CTC, sequence-level cross-entropy, from-scratch full-sum

1. INTRODUCTION

The recent sequence-to-sequence (seq2seq) acoustic models allow for from-scratch training within a unified optimization framework for automatic speech recognition (ASR). The common sequence-level cross-entropy training for both transducer based models with different label topologies [1, 2, 3], and attention based encoder-decoder model [4] do not necessarily require an initial alignment between the speech signal features and the output labels. The simplest case among the mentioned approaches is Connectionist Temporal Classification (CTC). The presence of the blank in the CTC topology gives more freedom to the alignment model which has less convergence problems during full-sum training, demanding at the same time less computation due to the independence assumption at each time step. However, the peaky behavior caused by the limited label emissions can also affect the quality of the alignment, which can shift with respect to the evidence in the input.

The typical training pipelines deployed for standard hybrid hidden Markov model deep neural network (HMM-DNN) [5] on the other hand requires the bootstrapping of a separate context-dependent Gaussian Mixture Model (GMM). The alignment taken from the GMM system is generally known to be a good starting point for subsequent ASR tasks or data segmentation, despite its low ASR accuracy. The hybrid model is generally trained with frame-wise cross-entropy (Fw-CE) using GMM alignment with a multi-stage complex pipeline consisting of various training criteria.

Previous work on single stage from-scratch systems considered Lattice-Free Maximum Mutual Information criterion (LF-MMI) [6, 7] and its further extension [8], using both CTC and HMM. However, the training can require high computations with still not fully competitive ASR results. Another possible from-scratch training solution is to use the maximum likelihood (ML) criterion, which is shown to have convergence problems that require complex training schedules and even the inclusion of additional losses [9]. Moreover, the quality of the Viterbi alignment taken from such from-scratch ML trained HMM based model even with an extended multi-task loss is still shown to lead to an inferior performance compared to a GMM based alignment, when used for further Fw-CE training in hybrid approach [10].

In this work, we compare the CTC and HMM topologies in the context of from-scratch full-sum training. In addition to the standard hybrid HMM, we propose a discriminative HMM modeling as a more direct contrast to CTC models. We then address the convergence issue from point of view of difficulty of alignment modeling, and propose several approaches in that regard. Together with the common subsampling, we propose both minimum label duration and prior-knowledge-based probability approximation. Experiments on both Switchboard and LibriSpeech copora show that these approaches not only benefit HMM models, but also improve CTC models. Besides ASR accuracy, we also analyze their capability of generating good alignment. With all these, we hope to initiate the path towards a simple from-scratch trainable system producing both good word error rate and alignment.

2. MODELS

Let X and W denote the acoustic feature sequence and corresponding word sequence of a speech utterance, respectively. Let $h_1^T = E(X)$ denote the common encoder output which
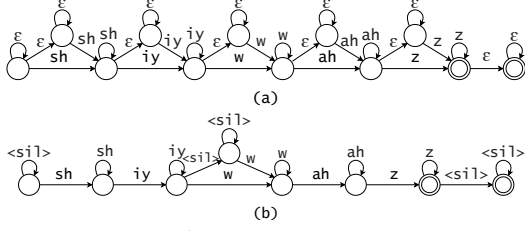

⋆ Denotes equal contribution

Fig. 1: FSA example of (a) CTC and (b) HMM topologies for the phoneme sequence of the utterance "*she was*"

Fig. 2: Modified FSA of (a) CTC and (b) HMM topologies with MinDur (k=2) for the first word "*she*" in Fig. 1

transforms X into high-level representations with optional subsampling. In the following, we consider the two most common label topologies for time-synchronous 0-order models, namely CTC and HMM, in the context of from-scratch full-sum training. With a major focus on phonetic-based acoustic modeling (AM), we provide detailed formulations in terms of models, training and decoding. We denote a_1^S (usually $T > S$) as the phonetic output label sequence of W and omit pronunciation variants for simplicity.

2.1. CTC

The standard CTC topology [1] introduces blank ϵ-augmented alignment sequences y_1^T to align a_1^S to the T frames. An example of possible alignments in the form of finite state automaton (FSA) is shown in Fig. 1. Each a_s has to occur at least once in y_1^T and optional ϵ can occur at any label segment boundaries. For label repetitions in a_1^S, at least one ϵ has to occur between the two labels in y_1^T. This allows a unique mapping of y_1^T to a_1^S by removing all label loops and ϵ. Thus, the posterior of the output label sequence can be written as:

$$P(a_1^S|X) = \sum_{y_1^T:a_1^S} P(y_1^T|h_1^T) = \sum_{y_1^T:a_1^S} \prod_{t=1}^T P(y_t|h_t) \quad (1)$$

Based on Eq. (1), the full-sum training for CTC models can be carried out from-scratch with the following loss:

$$\mathcal{L} = -\log\ P(a_1^S|X) \quad (2)$$

The final best word sequence $\hat{W}$ can be obtained via Viterbi decoding as:

$$X \to \hat{W} = \underset{W}{\arg\max} \left[P_{\text{LM}}^\lambda(W) \max_{y_1^T:a_1^S:W} \prod_{t=1}^T \frac{P(y_t|h_t)}{P_{\text{prior}}^\alpha(y_t)} \right] \quad (3)$$

Here an optional P_{prior} with scale α can be included, which can be obtained by marginalizing the framewise posterior on the training data. For $\alpha = 0$, this corresponds to a simple log-linear combination of the AM and language model (LM) with scale λ.

2.2. HMM

In this work, we mainly consider the HMM-0-1 topology, where only forward and loop transitions are allowed. As shown by the FSA in Fig. 1, the alignment sequence in this case contains only a_s with optional loops, while optional silence is also introduced at word boundaries. To be consistent
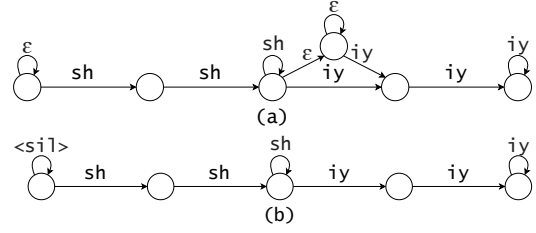
with standard HMM notations, here we use the state sequence s_1^T to represent alignments for a_1^S as $y_t = a_{s_t}$, which is also necessary to distinguish label repetitions in a_1^S. Note that theoretically there is no deterministic bijective mapping between the HMM and CTC topologies for any alignments.

We study the HMM topology with both discriminative and generative modeling approaches. We adopt the Markov assumption for state dependency and further simplify that with a pooled transition $\mathcal{T}(s_t, s_{t-1})$, which only has 4 values for loop/forward of speech/silence.

2.2.1. Posterior HMM

We call the discriminative approach as posterior HMM (**P-HMM**), where the output sequence posterior is defined as:

$$P(a_1^S|X) = \sum_{s_1^T} P(a_1^S, s_1^T|h_1^T)$$

$$= \sum_{s_1^T:a_1^S} \prod_{t=1}^T P(a_{s_t}, s_t|s_{t-1}, h_t)$$

$$= \sum_{s_1^T:a_1^S} \prod_{t=1}^T P(a_{s_t}|s_{t-1}^t, h_t) P(s_t|s_{t-1}, h_t)$$

$$= \sum_{s_1^T:a_1^S} \prod_{t=1}^T P(a_{s_t}|h_t) \mathcal{T}(s_t, s_{t-1}) \quad (4)$$

The full-sum training of P-HMM can then be carried out from scratch based on Eq. (2) and Eq. (4), where we can further apply exponential scales on the two probability terms in Eq. (4).

Similar as Eq. (3), the final best word sequence $\hat{W}$ can be obtained via Viterbi decoding as:

$$X \to \hat{W} = \quad (5)$$

$$\underset{W}{\arg\max} \left[P_{\text{LM}}^\lambda(W) \max_{s_1^T:a_1^S:W} \prod_{t=1}^T \frac{P(a_{s_t}|h_t)}{P_{\text{prior}}^\alpha(a_{s_t})} \mathcal{T}^\beta(s_t, s_{t-1}) \right]$$

where the optional P_{prior} can still be applied.

To explore the strong similarity with CTC models, we further simplify the P-HMM with a fixed $\mathcal{T}(s_t, s_{t-1}) = 0.5$ for all types of transitions. This allows to directly drop the term $\mathcal{T}(s_t, s_{t-1})$ in both training and decoding, which yields the correlation between Eq. (1) vs. Eq. (4) and Eq. (3) vs. Eq. (5). We call this simplified version as **P-HMM-S**.

2.2.2. Hybrid HMM

The generative modeling approach is the classical hybrid HMM (**H-HMM**), which defines

$$P(X|a_1^S) = P(h_1^T|a_1^S) = \sum_{s_1^T} P(h_1^T, s_1^T|a_1^S)$$

$$= \sum_{s_1^T:a_1^S} \prod_{t=1}^{T} P(h_t|a_{s_t})P(s_t|s_{t-1})$$

$$= \sum_{s_1^T:a_1^S} \prod_{t=1}^{T} \frac{P(a_{s_t}|h_t)}{P_{\text{prior}}(a_{s_t})} \mathcal{T}(s_t, s_{t-1}) \qquad (6)$$

The from-scratch full-sum training of H-HMM can then be carried out by minimizing the loss

$$\mathcal{L} = -\log P(X|a_1^S) \qquad (7)$$

where we can further apply exponential scales on the three probability terms in Eq. (6). Despite the tuning effort, we find that an optimal scaling is essential for both convergence and final best performance.

The decoding of the best word sequence using H-HMM is the same as Eq. (5). Note that the major difference to P-HMM is the inclusion of P_{prior} in training, which effectively results from the discriminative and generative modeling nature.

2.3. Label Units

Following [11], we apply end-of-word (EOW) augmentation to the phoneme sets by default. For the granularity of a_1^S, we consider both single-state and three-states structure for each phoneme. The latter is commonly applied in the classical hybrid HMM approach to account for the start, middle, and end of a phoneme.

3. CONVERGENCE AND ALIGNMENT MODELING

The full-sum training with either CTC or HMM topology is effectively an Expectation Maximization (EM)-style alignment modeling problem. The general procedure relies on the calculation of the occupation probability q_t, known as soft-alignment, at each EM iteration. The computation of the soft-alignment can be done via the Baum-Welch forward-backward algorithm. For a given input h_1^T, and a label sequence a_1^S, the quantity $q_t(a_s|h_1^T, a_1^S, \theta)$ represents the probability mass of all alignment paths going through the state s at time frame t, according to the parameters θ. In practice, the quantity q_t is a normalized probability distribution over the set of all labels at time frame t. With the maximum approximation, the label of aligned state in the best path has probability one. The latter alignment is known as hard-alignment or Viterbi forced alignment.

Ideally, the model in training should converge to more prominent alignment paths and assign higher probabilities to them. The difficulty of such alignment modeling can be affected by the number of possible alignment paths, which is correlated with the relation between T and S (usually $T \gg$

S). This is then directly related to the convergence issue of neural network (NN) training (generally non-convex optimization), especially from scratch.

With blank allowed almost anywhere, the CTC topology reveals an easy optimum for NN training to assign most probability masses to alignments with most blank frames. This allows an easy convergence for CTC models, but also leads to the common peaky behavior [12]. Thus, CTC models usually is not able to produce Viterbi alignment with high correspondence to the actual speech signal, where the ambiguity of blank even enlarges the difficulty [8]. To solve this, one might need to force certain label loops and force blank to cover silence only, which eventually forces the CTC topology towards the HMM topology. On the other hand, the HMM topology appears to be more restricted and thus, may result in more difficulty in convergence [9]. Here we propose to solve the convergence issue by reducing the difficulty of alignment modeling:

- **subsampling (SS)** is widely applied in modern ASR systems to reduce computation and memory cost. By reducing T, subsampling effectively also reduce the number of possible alignment paths and therefore, can largely simplify alignment modeling for an easier convergence (extreme case: $T \approx S$).

- **minimum duration (MinDur)** of speech label forces each a_s to occur at least k times in the alignment. This shares the same spirit as subsampling to reduce the number of possible alignment paths and thus, can also ease the convergence issue. In addition, it also forces the model to be less peaky, which might lead to more reasonable alignment paths. The minimum duration, as shown in Fig. 2, can be simply applied by modifying the FSA structure for full-sum training.

- **probability approximation** (P-**approx**) uses prior knowledge to estimate those transition and prior probabilities in P-HMM or H-HMM. These values are then kept fixed during training to guide the model towards more reasonable alignment paths. For our phonetic-based AM, we make use of the fact that a phoneme is commonly of 80ms long on average, which reveals the speech forward/loop probabilities as $\mathcal{T}(s_t = s_{t-1}) = 7/8$ under a 10ms-frame scenario. Together with the transcription of acoustic training data, we can also derive the prior probability of each phoneme. Then by regarding the remaining audio length of each utterance all to be silence, we can also obtain the silence prior probability. Finally, by regarding these silence length of each utterance as equally from sentence begin/end, we obtain the silence transition probability for $\mathcal{T}$.

These methods can also be combined with minor adjustment. We also apply subsampling and minimum duration to CTC models for fair comparison.

Table 1: Performance of our models in terms of word error rates (WER) on Hub5'0{0,1} using a 4-gram language model. We show the effect of using SS, MinDur and P-approx (for $\mathcal{T}$ and/or P_{prior}). We consider single-state or three-states phoneme with EOW. We denote $\times$ as not converged model.

Model	States	SS	MinDur	P-approx	WER [%] Hub5'00	Hub5'01
CTC	1				17.6	16.0
		✓			**13.7**	**13.1**
			✓		14.6	14.0
P-HMM-S					$\times$	
		✓			14.5	13.9
			✓		15.0	-
P-HMM				✓	17.5	-
		✓		✓	14.3	13.9
	3			✓	16.2	-
H-HMM	1			✓	24.0	-
	3			✓	13.9	13.5

Table 2: Similar experiments for the most promising models presented in Table 1, trained with LibriSpeech 960 hours, and evaluated on all four test and dev datasets.

Model	States	SS	MinDur	P-approx	dev clean	dev other	test clean	test other
CTC	1				4.6	11.3	5.0	12.3
		✓			**3.2**	**8.0**	**3.6**	**8.5**
			✓		3.7	8.9	4.1	9.6
P-HMM-S					$\times$			
		✓			3.4	8.5	3.8	9.4
			✓		3.7	9.1	-	
P-HMM		✓		✓	3.5	8.9	-	
H-HMM	3			✓	3.6	9.1	3.9	9.5

4. EXPERIMENTS

4.1. Setting

We conduct our experiments on two separate corpora, 960 hours LibriSpeech (LBS) [13], and 300 hours Switchboard (SWB) (LDC97S62) [14]. The evaluations are done on all four dev and test sets for LBS, as well as on SWB and Call-Home (CH) subsets of Hub5'00 (LDC2002S09) and the three SWB subsets of Hub5'01 (LDC2002S13). We use the official lexicon of LBS and SWB, where we unify the stressed phonemes for the former [15]. EOW phonemes are applied in all cases.

All acoustic models use a 6×512 BLSTM encoder and the optional subsampling of factor 4 is done via 2 max-pooling layers, in the middle. The input features to the encoder are (LBS: 50 and SWB: 40) dimensional Gammatone Filterbank features [21], extracted from 25 milliseconds (ms) window with 10ms shift. We train our models for (LBS: 25 and SWB: 30) epochs, combining two learning rate (LR) scheduling. An initial one-cycle learning rate (OCLR) policy [22] covers 90% of the total steps with linear increase up to peak LR of around $3e^{-3}$ and consequent linear decrease, following [15]. We continue the remaining steps with a constant minimum LR of $1e^{-5}$. The Adam optimizer with Nesterov momentum [23] is used. As regularization techniques we use 10% dropout [24] and L_2 weight decay with a scale of $1e^{-4}$. Training of P-HMM and H-HMM includes also the choice of

Table 3: Evaluation of our best model for SWB together with some results from the literature. WE denote monophone and diphone as {mono,di}P, and specify the number of states ({1,3}-S), with optional EOW augmentation.

Work	Model	Label	From-Scratch	Criterion	LM	WER [%] SWB	CH	$\sum$
[16]	H-HMM	monoP-3S	no	Fw-CE	4-gram	9.8	**18.4**	14.1
[17]	HMM	diP-2S	yes	LF-MMI	RNN	9.8	19.3	-
[18]	CTC	1P-1S		ML	4-gram	13.9	24.7	
This Work	CTC	EOW-1P-1S				**8.8**	18.5	**13.7**
This Work	H-HMM	EOW-1P-3S				9.3	18.5	13.9

Table 4: Comparison of our models with other approaches in the literature for LBS 960h, with evaluation on all dev and test sets. We considered also triphone CART label (triP) and wordpieces (WP).

Work	Model	Label	From-Scratch	Criterion	LM	dev clean	dev other	test clean	test other
[19]	H-HMM	triP-3S	no	Fw-CE	4-gram	4.0	9.6	4.4	10.0
[8]	HMM	WP	yes	bLF-MMI	4-gram	3.9	9.7	-	
[20]	Wav2L	1P-1S			KenLM	-		7.2	-
This Work	CTC	EOW-1P-1S		ML	4-gram	**3.2**	**8.0**	**3.6**	**8.5**
This Work	P-HMM-S	1P-1S				3.4	8.5	3.8	9.4

the scales α, β, and γ for the state prior, transition and label posterior, respectively. The optimal setting for minimum duration is (LBS: 3, SWB: 4) and (LBS: 4, SWB: 5) for CTC and HMM, respectively.

Training of our models rely on the combination of RASR and RETURNN toolkits [25, 26], where the weighted FSA for each utterance is constructed within the former, and the NN training and the CUDA based BW computations are done by the latter. The decoding makes use of time synchronous lexical prefix trees search within RASR [27, 28]. By default we use word-level 4-gram LM in all cases.

4.2. ASR Accuracy

We present our results for different modeling approaches in Tables 1 and 2, and compare our best model with the literature in Tables 3 and 4. For fair comparison, we consider only from-scratch 0-order models from the literature, when possible.

We show that the application of either subsampling or minimum duration to CTC improves the accuracy on both corpora, with the former performing the best among all proposed models. Similarly, we can see that P-HMM-S with subsampling performs better than with minimum duration, on both corpora. The HMM topology without any additional method does not converge for any of the two copora. On SWB task we observe that by adding the transition probability the model converges to 17.5% WER on Hub5'00. This can be improved by either switching to a three-states model or by application of subsampling, reaching 16.2% and 14.3%, respectively. The combination of subsampling and fixed transition helped the P-HMM also on LBS task, obtaining 3.5% and 8.9% on dev-clean and dev-other, respectively. The H-HMM with its 13.9% and 13.5% on the two Hub5 datasets, show only relative 1% and 2% degradation compared to the

Table 5: The calculation of TSE on SWB train and Hub5'00 dev-sub with respect to a GMM monophone alignment for our proposed models and a standard BLSTM hybrid HMM trained with Fw-CE using a tandem based alignment.

Plot	Model	States	SS	MinDur	P-approx	TSE [ms] train	TSE [ms] dev-sub
-	H-HMM-CE	3				41	39
a.1	CTC					86	67
a.2	CTC		✓			67	56
a.3	CTC	1		✓		**52**	**49**
a.4	P-HMM-S		✓			68	56
a.5	P-HMM-S			✓		72	68
a.6	P-HMM	3			✓	63	57
a.7	H-HMM				✓	106	89

Table 6: The effect on convergence during training of the state prior, label posterior, and transition scales, α, γ, and β, respectively. All experiments are for SWB300h and evaluated on Hub5'00 with 4-gram LM.

Model	Approach	α	γ	β	WER [%]
P-HMM	P-approx	N/A	1.0	1.0	×
			0.3	0.1	17.5
	P-approx+SS		1.0	0.1	15.0
				0.01	14.4
H-HMM	P-approx	1.0	1.0	1.0	×
		0.5			20.3
		0.3	0.3	0.1	17.5
		0.1			15.1
				0.3	14.7

CTC with subsampling. However, this gap becomes larger on LBS task.

Our best from-scratch models with CTC and HMM topologies on both SWB and LBS obtained competitive results compared to other approaches in the literature. Our H-HMM, as reported in Table 3, obtains slightly better performance than the hybrid monophone trained with a tandem alignment [29]. On LBS task included in Table 4, our P-HMM-S outperforms a triphone hybrid trained with triphone GMM alignment. CTC with subsampling again outperforms all models.

4.3. Analysis

4.3.1. Alignment Plots

Whether a high performance ASR model is also able to provide a reasonable alignment with accurate time stamp for subsequent ASR or segmentation tasks is still an open question. Therefore, in addition to the evaluation of the ASR accuracy of our models reported in Tables 1 and 2, we carried out further investigations on the quality of the alignment taken from our most promising models. For this purpose, we used SWB 300h train and a subset of Hub5'00, which we denote as **dev-sub**, where all segments containing out-of-vocabulary words with respect to the official SWB lexicon are excluded.

Since the occupation probability is computed using the model parameters at each iteration, we can assume that a soft-alignment with concentrated probability mass around the most prominent path can indicate good model convergence. This path eventually is very near or almost the same as what

we obtain by doing forced alignment. This supposition is confirmed by the plots (a.n) and (b.n) with n $\in \{1, \cdots, 7\}$, for Viterbi and BW alignments, respectively, as shown in Fig. 3. Generally speaking, we believe that the evaluation of the quality of an alignment is yet an undecidable issue. There is no ground-truth and therefore no defined evaluation metric for measuring the quality of an alignment. This becomes even a more important problem when we compare the alignment across different topologies. We chose as reference a GMM monophone alignment, and evaluated our Viterbi alignments by considering the mean absolute distance (in milliseconds) of word start and end-positions against the reference alignment. Averaging such distance over all words gives us the time-stamp-error (TSE) metric, as used in [8]. Note that silence is not counted for TSE, which also makes the evaluation straightforward for CTC models. An overview of different TSE values is shown in Table 5. Since the alignment learned by a BLSTM base AM can have further shifting due to the NN encoder, we also considered the TSE of a standard H-HMM trained on a tandem system alignment with Fw-CE.

In Section 4.2, we showed that subsampling helps to counteract the convergence problem. The effectiveness of this method is confirmed also for the alignment quality, where CTC with subsampling has a better TSE. This is the same number of 10 ms frames on both train and dev-sub of our P-HMM-S trained with subsampling. Even though the ASR accuracy of CTC with minimum duration is worse than CTC with subsampling, we can see that the former produces a better alignment. An aspect that is valid also for the case of P-HMM and H-HMM. When we switch to the HMM topology, both minimum duration and the choice of the three-states model introduce larger constraint on the phoneme duration, due to the presence of the additional states in the FSA. This causes a reduction in the silence duration in the alignment which leads to a larger shift and therefore higher TSE. This can be seen when one compares the plots a.4 and a.5 of P-HMM-S with subsampling and minimum duration, respectively. The duration of the silence is further reduced in the alignment when we subtract the state prior during training, as a result of a higher penalty on silence. This can be seen by comparing the plot a.6 of P-HMM with the plot a.7 of H-HMM. Similarly to the CTC case, we have a three-states H-HMM that produces an alignment with higher TSE but has a better ASR accuracy compared to a three-states P-HMM. The high silence probability in the latter may force the time synchronous beam search process to wrongly prune away correct speech states. However, the absence of the silence in the alignment taken from the H-HMM has the opposite effect of causing a large shift with respect to the evidence in the signal.

4.3.2. Effect of the Scales for HMM

The modeling approaches described in Section 2.2 make use of optional scales during both training and decoding. This
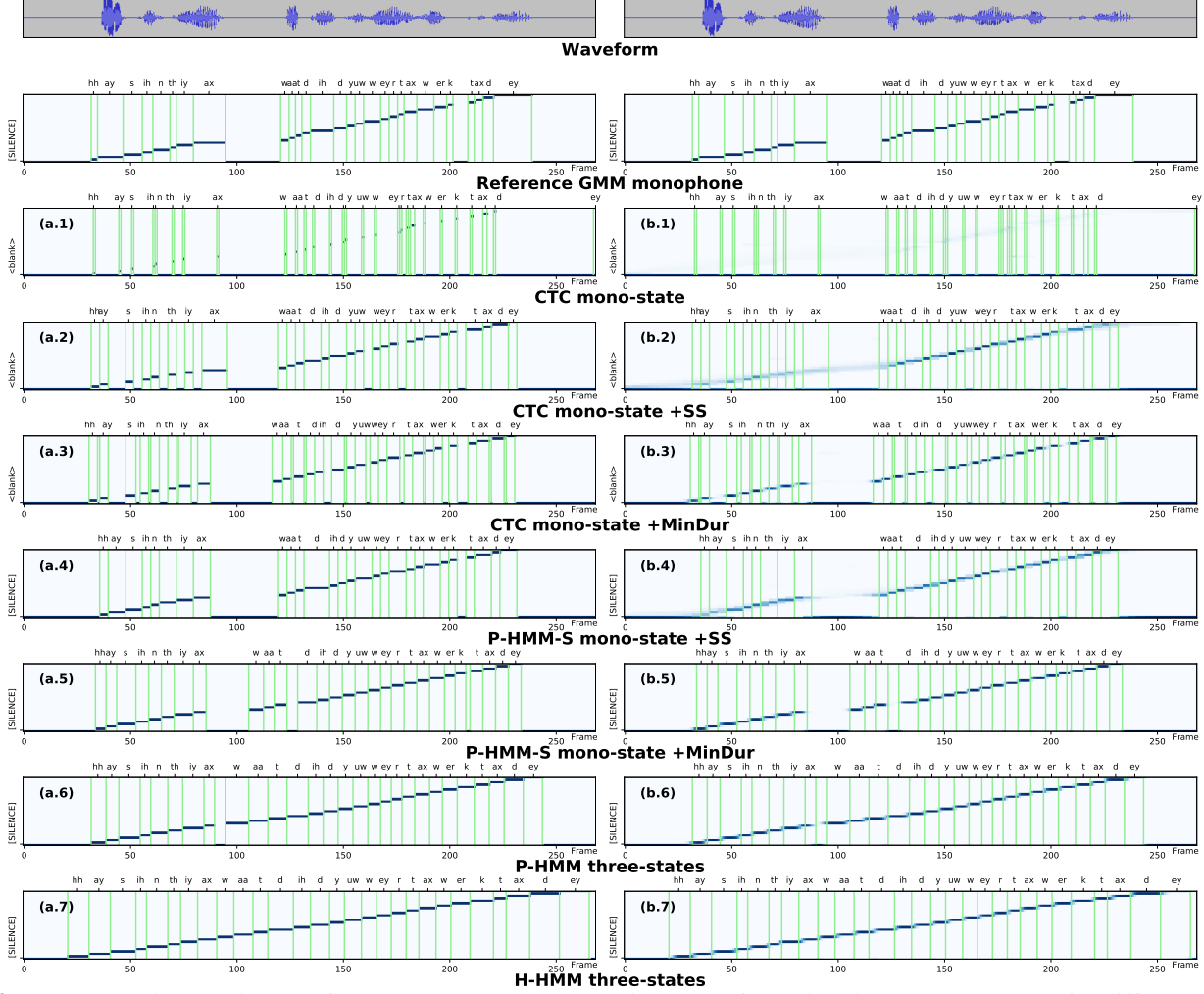
Fig. 3: The comparison against a reference GMM monophone alignment of (a) Viterbi and (b) BW paths for different HMM and CTC topologies using our proposed methods for full-sum training from-scratch. The horizontal lines indicates the phoneme boundaries.

means one can control the contribution of each factor via the scales with consequent variations of ASR accuracy. We observed that for a given set of fixed scales used during full-sum training, the effect of tuning of the decoding scales on the WER fluctuations are small, similarly to our common experience with Viterbi trained models. However, the choice of scales for the full-sum from-scratch training of HMM topology, as shown in Table 6, turns out to be of a significant importance.

5. CONCLUSION

In this work, we performed an in-depth study on from-scratch full-sum training for both CTC and HMM topologies, where we proposed the posterior HMM as a middle ground between standard CTC and hybrid HMM. We discussed different methods to counteract the convergence issue during from-scratch full-sum training. We showed that by addressing the problem from point of view of difficulty of the alignment modeling, we can improve not only the ASR accuracy, but also the time-stamp-error of our models with respect to a

reference GMM alignment. Our proposed methods, along with commonly used subsampling, include also novel use of minimum duration and prior-knowledge-based probability approximations. Our single-stage trained models show competitive word error rates on Switchboard and LibriSpeech tasks compared to other approaches in the literature. For both topologies, we also showed that higher ASR accuracy does not always lead to a better alignment quality. We foresee the investigation of a modeling approach that fulfills both requirements as the main future direction of our work.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. of Intern. Conf. on Machine Learning (ICML)*, 2006.

[2] Alex Graves, "Sequence Transduction with Recurrent Neural Networks," *Proc. of Intern. Conf. on Machine Learning (ICML)*, 2012.

[3] Hasim Sak, Matt Shannon, Kanishka Rao, and Françoise Beaufays, "Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence to Sequence Mapping," in *Proc. of Interspeech*, 2017.

[4] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based Models for Speech Recognition," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.

[5] Herve A Bourlard and Nelson Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, vol. 247, Springer Science & Business Media, 2012.

[6] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely Sequence-trained Neural Networks for ASR based on Lattice-free MMI," in *Proc. of Interspeech*, 2016.

[7] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur, "End-to-end Speech Recognition Using Lattice-free MMI.," in *Proc. of Interspeech*, 2018.

[8] Xiaohui Zhang, Vimal Manohar, David Zhang, Frank Zhang, Yangyang Shi, Nayan Singhal, Julian Chan, Fuchun Peng, Yatharth Saraf, and Mike Seltzer, "On Lattice-free Boosted MMI Training of HMM and CTC-based Full-context ASR Models," in *IEEE ASRU*, 2021.

[9] Albert Zeyer, Eugen Beck, Ralf Schlüter, and Hermann Ney, "CTC in the Context of Generalized Full-sum HMM Training," in *Proc. of Interspeech*, 2017.

[10] Tina Raissi, Eugen Beck, Ralf Schlüter, and Hermann Ney, "Towards Consistent Hybrid HMM Acoustic Modeling," *arXiv:2104.02387*, 2021.

[11] Wei Zhou, Simon Berger, Ralf Schlüter, and Hermann Ney, "Phoneme based Neural Transducer for Large Vocabulary Speech Recognition," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2021.

[12] Albert Zeyer, Ralf Schlüter, and Hermann Ney, "Why does CTC result in Peaky Behavior?," May 2021, arXiv:2105.14849.

[13] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: An asr corpus based on public domain audio books," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2015.

[14] John J Godfrey, Edward C Holliman, and Jane McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 1992.

[15] Wei Zhou, Wilfried Michel, Ralf Schlüter, and Hermann Ney, "Efficient Training of Neural Transducer for Speech Recognition," 2022.

[16] Tina Raissi, Eugen Beck, Ralf Schlüter, and Hermann Ney, "Improving Factored Hybrid HMM Acoustic Modeling without State Tying," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2022.

[17] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur, "Flat-start Single-stage Discriminatively Trained HMM-based Models for ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1949–1961, 2018.

[18] Kartik Audhkhasi, Bhuvana Ramabhadran, George Saon, Michael Picheny, and David Nahamoo, "Direct Acoustics-to-word Models for English Conversational Speech Recognition," in *Proc. of Interspeech*, 2017.

[19] Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "RWTH ASR Systems for LibriSpeech: Hybrid vs Attention," *Proc. of Interspeech*, 2019.

[20] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve, "Wav2letter: An End-to-end Convnet-based Speech Recognition System," *arXiv:1609.03193*, 2016.

[21] Ralf Schlüter, Ilja Bezrukov, Hermann Wagner, and Hermann Ney, "Gammatone Features and Feature Combination for Large Vocabulary Speech Recognition," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2007.

[22] Leslie N Smith and Nicholay Topin, "Super-convergence: Very Fast Training of Neural Networks using Large Learning Rates," in *Artificial intelligence and machine learning for multi-domain operations applications*. SPIE, 2019, vol. 11006, pp. 369–386.

[23] Timothy Dozat, "Incorporating Nesterov Momentum into Adam," 2016.

[24] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of machine learning research*, 2014.

[25] Simon Wiesler, Alexander Richard, Pavel Golik, Ralf Schlüter, and Hermann Ney, "RASR/NN: The RWTH Neural Network Toolkit for Speech Recognition," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2014.

[26] Albert Zeyer, Tamer Alkhouli, and Hermann Ney, "RE-TURNN as a Generic Flexible Neural Toolkit with Application to Translation and Speech Recognition," in *demo at ACL*, 2018.

[27] Hermann Ney, Reinhold Haeb-Umbach, B-H Tran, and Martin Oerder, "Improvements in Beam Search for 10000-word Continuous Speech Recognition," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 1992.

[28] Hermann Ney and Stefan Ortmanns, "Dynamic Programming Search for Continuous Speech Recognition," *Signal Processing Magazine*, 1999.

[29] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, "Speaker Adaptive Joint Training of Gaussian Mixture Models and Bottleneck Features," in *IEEE ASRU*, 2015.