



Investigating the Effect of Label Topology and Training Criterion on ASR Performance and Alignment Quality

Tina Raissi¹, Christoph Lüscher^{*2}, Simon Berger^{*1,2}, Ralf Schlüter^{1,2}, Hermann Ney^{1,2}

¹Machine Learning and Human Language Technology Group, RWTH Aachen University, Germany
²AppTek GmbH, Germany

* Denotes equal contribution

{raissi,schlueter}@cs.rwth-aachen.de

Abstract

The ongoing research scenario for automatic speech recognition (ASR) envisions a clear division between end-to-end approaches and classic modular systems. Even though a high-level comparison between the two approaches in terms of their requirements and (dis)advantages is commonly addressed, a closer comparison under similar conditions is not readily available in the literature. In this work, we present a comparison focused on the label topology and training criterion. We compare two discriminative alignment models with hidden Markov model (HMM) and connectionist temporal classification topology, and two first-order label context ASR models utilizing factored HMM and strictly monotonic recurrent neural network transducer, respectively. We use different measurements for the evaluation of the alignment quality, and compare word error rate and real time factor of our best systems. Experiments are conducted on the LibriSpeech 960h and Switchboard 300h tasks.

Index Terms: speech recognition, acoustic modeling, finite state transducers, alignment quality

1. Introduction

According to the latest version of the survey *Progress and Prospects for Spoken Language Technology*, there is a general consensus that by the year 2025 the majority of automatic speech recognition (ASR) systems have completely abandoned the hidden Markov model (HMM) paradigm [1]. Even though rich relations exist between HMM and current approaches often termed as end-to-end [2]. In addition to the first-order Markov assumption, HMM builds upon a conditional independence assumption. According to this assumption, when conditioned on a state, the aligned frames within that segment are independent of each other, as well as of all the frames generated by the other states. Nearly all HMM based alignment systems used in literature are trained by maximizing the likelihood of the training data with respect to a set of Gaussian mixture model parameters. A generative model such as GMM needs to cover for all possible observation sequences and therefore requires that the input sequence is restricted to a local representation with no access to wider ranges of dependencies. In the years following the success of large vocabulary ASR with hybrid-HMM, substantial research efforts have been directed towards addressing the inherent simplifying assumptions in hidden Markov approach. This has been done by considering a wider range of dependencies, applying global normalization, and performing sequence discriminative training [3, 4, 5, 6]. In the recent years, there has been exploration into the simplification of the standard HMM pipeline by eliminating GMM and HMM state-typing [7, 8]. Moreover, a simulation study verifying the effect of different simplifying assumptions of classic HMM has notably underlined the significant influence of the aforementioned

output independence assumption on the ASR performance [9]. The recent encoder architectures are equipped with the self-attention mechanism and recurrent layers which have the capability to capture global dependencies of the input. One question to consider is whether utilizing downsampling over the time axis within these long-context neural networks allows the creation of an input representation that is more in line with the HMM independence assumption.

The current research scenario seems to have a preference towards the end-to-end approaches, primarily due to their simplicity and the widespread accessibility of resources and frameworks. However, when comparing the ASR performance of these approaches with classic HMM-based systems, various issues arise: differences in speech representation and frame shifts on the front-end side, as well as variations in label units, amount of data, use of external resources such as lexicon, alignment, and language model [10, 11].

In this work, we propose a comparison between two pairs of time synchronous models, by fixing all the conditions mentioned above, and considering different label topologies and training criteria. The specific choice of the models derives from a more high-level comparison between a purely discriminative sequence-to-sequence approach and a factored hybrid HMM system. The pipeline for both systems starts with training a zero-order label context discriminative alignment model from-scratch. Subsequently, after conducting forced alignment, a first-order label context model is trained for each system. As a further investigation step, we analyze the quality of the alignments for each topology using different measurements, and show improvements with respect to both word error rate (WER) and real time factor (RTF) for factored hybrid HMM system with a 40ms frame shift. All experiments are verified on both LibriSpeech 960h and Switchboard 300h.

2. Modeling Approach

For an acoustic feature sequence X of length T' , and its corresponding word sequence W of length N , let $h_1^T = Enc(X)$ denote the output of a neural encoder. We consider $T = \lceil T'/m \rceil$ for $m \in \{1, 4\}$, where $m = 4$ realizes the common downsampling technique used within the neural encoder architectures. The inference rule for any ASR task is Bayes decision rule [12], which maximizes the a-posteriori probability of a word sequence W given the input sequence, as described in Eq. (1). In the classic approach, an equivalent reformulation via Bayes identity in terms of separate acoustic and language models is carried out. The resulting generative approach having distinct parameters θ_{AM} and θ_{LM} , respectively, is shown in Eq. (2).

$$X \rightarrow \hat{W}(X) = \underset{W}{\operatorname{argmax}} \{P_{\theta}(W|X)\} \quad (1)$$

$$= \underset{W}{\operatorname{argmax}} \{P_{\theta_{AM}}(X|W) \cdot P_{\theta_{LM}}(W)\} \quad (2)$$

2.1. Unifying Principles

In order to verify the effect of the label topology and the training criterion across different models, we fix some of the basic conditions. The underlying speech representation adheres to the classic HCLG [13] composition unfolded over time, i.e. an alignment finite state acceptor (FSA) \mathcal{A} that defines all valid alignment paths between the input speech and the output label sequence. Independent from the modeling approach, the word sequence is generally mapped to a sequence of sub-word units. Let the sequence ϕ of length M be the phonetic output label sequence of W , and disregard pronunciation variants. We augment each phoneme with its neighboring left and right phonemes for the classic triphone structure of the C composition in HCLG. Unfolding the result of the composition with a specific label topology over time is equivalent to a marginalization over a hidden alignment variable. In practice it gives the FSA \mathcal{A} that contains all the allowed paths within a certain topology. The training criteria used for our models belong to the cross-entropy based family of training objectives. The criterion considered on the sequence level uses the sum over \mathcal{A} , and on the frame level restricts the loss to a single best path. With single state phonemes as label units, we explore different label topologies and probability distribution models for the frame-wise acoustic scores, as later explained in Sec. 2.3. The label set for all topologies consists of the phonemes from the lexicon with the end-of-word (EOW) distinction [14]. For CTC and transducer an additional blank token is included and only HMM uses an explicit silence token. We use a downsampling of factor four on the ten milliseconds (ms) input features for a total frame shift of 40ms.

2.2. Mapping of the Alignment Path

We showed how a word sequence can be mapped to its alignment FSA \mathcal{A} . For the inverse mapping within our proposed notation, we introduce two auxiliary functions: (1) for the blank-based topologies, denote Y as a blank-augmented alignment sequence of length T , we use a function \mathcal{B} which eliminates all blanks for an alignment path within \mathcal{A} and returns the corresponding phoneme sequence, and (2) for the HMM topology, by using the hidden state sequence S of length T , we denote $a_{n_{s_t}}$ as a function that takes as input the identity of the aligned state at time frame t within a phoneme of the word at position n and returns the phoneme label.

2.3. Conceptual Framework

Without using the WFST formalism, we describe each approach via a three-fold conceptual framework. Denote $\mathcal{F} = \{\Sigma, \mathcal{Q}, \mathcal{C}\}$ to be a tuple consisting of a label topology Σ , a frame-wise probability distribution model \mathcal{Q} , and a training criterion \mathcal{C} .

- **Label Topology Σ** : describes a set of rules related to the emission of a label out of the set of labels, at each time step. A label topology together with the label unit determine the set of all allowed paths within the alignment lattice \mathcal{A} .

- **Frame-level probability distribution model \mathcal{Q}** : is obtained via application of model-specific assumptions for the decomposition of the sequence-level probability distributions of Eqs. (1) and (2) into frame-level probability models. The decomposition step generally follows a marginalization over a hidden alignment sequence variable. In the classic FST framework, \mathcal{Q} can be simplified to the type of acoustic weight associated with each arc. We model n -order label context acoustic models for $n \in \{0, 1\}$ with local normalization, i.e. with normalized probability distribution over the set of all labels at each time frame. At each time frame, we model posterior probabilities condi-

tioned on the encoder output for both orders and conditioned additionally on the previous acoustic label context for the first order models.

- **Training Criterion \mathcal{C}** : expresses the optimization objective. While all considered criteria can be viewed as a form of cross-entropy, there are differences caused by the specific modeling assumptions. For instance, we consider a sequence-level conditional likelihood with sum over all alignment paths, denoted as full-sum criterion, and a frame-wise cross-entropy or Viterbi training using the single best path.

2.4. Models

2.4.1. Zero-Order Label Context Alignment Models

We examine two discriminative alignment models that differ by label topology and frame level probability distribution. Both models are trained from scratch with conditional log-likelihood criterion. One model uses the classic CTC label topology with blank and label loop [15]. The second model, proposed previously as posterior-HMM (P-HMM) [16], consists in an HMM topology with loop and forward transitions only. We start with the marginalization of the acoustic model with the hidden alignment variable, and we further insert the label topology constraint and model-specific assumptions while decomposing the sequence-level posterior into frame-level probability distributions. The definition of \mathcal{Q} for CTC in Eq. (3) and for P-HMM in Eq. (4) differs by the transition probability $P(s_t|s_{t-1})$. Both models use a zero-order label posterior. However, the HMM uses an explicit transition probability $P(s_t|s_{t-1})$ normalized over loop and forward probability.

$$\sum_Y P(\phi, Y|h_1^T) = \sum_{y_1^T: \phi_1^M} \prod_{t=1}^T P(y_t|h_t) \quad (3)$$

$$\sum_S P(\phi, S|h_1^T) = \sum_{s_1^T: \phi_1^M} \prod_{t=1}^T P(a_{n_{s_t}}|h_t)P(s_t|s_{t-1}) \quad (4)$$

The transition probability is generally omitted in case of CTC and given implicitly by the alignment sequence, unless considered within the Auto Segmentation Criterion [17]. Similar assumptions can be done also for the simple variant of the P-HMM where loop and forward probabilities may be set to 0.5 for all labels, and therefore they may be dropped, being constant with respect to both the training criterion and the inference rule [16]. Training is carried out for both models with conditional log-likelihood described in Eq. (5) as follows:

$$\mathcal{L} = -\log P(\phi|h_1^T) \quad (5)$$

In practice, the two models learn the parameters θ for the inference rule in Eq. (1) via two different alignment FSAs \mathcal{A} and using two different probability distributions \mathcal{Q} . For P-HMM, the inference rule, shown in Eq. (6), uses the label and transition probabilities with scales α , and β , respectively, combined with an external language model (LM) scaled by λ . Moreover, a common internal language model (ILM) subtraction with scale γ is applied. Decoding for CTC follows a similar rule without using the transition model.

$$\operatorname{argmax}_W \left\{ P_{\text{LM}}^\lambda(W) \max_{s_1^T: \phi_1^M: W} \prod_{t=1}^T \frac{P^\alpha(a_{n_{s_t}}|h_t)}{P_{\text{ILM}}^\gamma(a_{n_{s_t}})} P^\beta(s_t|s_{t-1}) \right\} \quad (6)$$

2.4.2. First-Order Label Context ASR Models

The two first-order label context models are also built upon a common training criterion. However, in addition to different label topologies and frame level probability distributions, the models are designed to provide the acoustic parameters for different inference rules. Regarding the label topology Σ , we compare a strictly monotonic RNN-T topology (mRNN-T) [18, 19]

and a context-dependent HMM without state-tying presented as factored hybrid (FH) [20]. Both models use a frame-wise cross-entropy loss. We first discuss the \mathcal{Q} component and the decision rule for the factored hybrid HMM case. The starting point in this approach is shown in Eq. (2) and the model we consider provides θ_{AM} , i.e. the parameters of the acoustic model in the generative form. The formulation in Eq. (7) defines a generative first-order label context (diphone) model, consisting of two separate factors, together with a context-dependent diphone state prior.

$$\sum_S P(h_1^T, S|\phi) = \sum_{s_1^T: \phi_1^M} \prod_{t=1}^T P(h_t | a_{n_{s_t}}, a_{n_{s_t-1}}) P(s_t | s_{t-1})$$

$$= \sum_{s_1^T: \phi_1^M} \prod_{t=1}^T \frac{P(a_{n_{s_t}} | a_{n_{s_t-1}}, h_t) P(a_{n_{s_t-1}} | h_t) P(h_t)}{P_{\text{Prior}}(a_{n_{s_t}}, a_{n_{s_t-1}})} P(s_t | s_{t-1}) \quad (7)$$

The decision rule for FH in Eq. (8) uses both the language model and the prior as part of the decomposition and model definition.

$$\operatorname{argmax}_W \left\{ P_{\text{LM}}^\lambda(W) \max_{s_1^T: \phi_1^M: W} \prod_{t=1}^T \frac{P(a_{n_{s_t}}, a_{n_{s_t-1}} | h_t)}{P_{\text{Prior}}^\gamma(a_{n_{s_t}}, a_{n_{s_t-1}})} P^\beta(s_t | s_{t-1}) \right\} \quad (8)$$

The decomposition for the label posterior in mRNN-T includes only one of the factors from Eq. (7), due to the direct discriminative approach:

$$\sum_Y P(\phi, Y | h_1^T) = \sum_{y_1^T: \phi_1^M} \prod_{t=1}^T P(y_t | a_{n_{y_{t-1}}}, h_t) \quad (9)$$

Note that here we overload the a_n function to accept a blank augmented alignment label. The language model used in Eq. (10) should be considered as an external LM and we subtract the internal LM instead of the context-dependent label prior.

$$\operatorname{argmax}_W \left\{ P_{\text{LM}}^\lambda(W) \max_{y_1^T: \phi_1^M: W} \prod_{t=1}^T \frac{P(y_t | a_{n_{y_{t-1}}}, h_t)}{P_{\text{ILM}}^\gamma(y_t | a_{n_{y_{t-1}}})} \right\} \quad (10)$$

3. Experimental Setting

The experiments are conducted on 300h Switchboard-1 (SWB) Release 2 (LDC97S62) [21] and 960h LibriSpeech (LBS) [22]. The evaluations for the SWB task are performed on SWB and CallHome subsets of Hub5*00 (LDC2002S09) and three subsets of Hub5*01 (LDC2002S13). For LBS we report WERs on dev and test sets. For training we utilize the toolkit RETURNN [23]. Decoding of HMM based models use RASR for the core algorithms, and its recent extension for CTC and mRNN-T decoding [24, 25]. Our experimental workflow is managed by Sisyphus [26]. The speech signal is extracted from a 25ms window with a 10ms shift resulting in (SWB: 40, LBS: 50) dimensional Gammatone filterbank features [27]. SpecAugment is applied to all models [28]. All first-order models utilize existing setups for Conformer encoder [8, 29]. The model size of our 12-layers Conformer encoders is $\sim 75\text{M}$ parameters. All alignment models utilize a recurrent encoder following a similar setup to one of our previous works [30]. The recurrent encoder consists of 6 BLSTM layers with 512 nodes per direction, having $\sim 46\text{M}$ parameters. The decision to use a recurrent encoder with fewer parameters rather than the Conformer encoder for the alignment model stems from earlier experimental results indicating comparable performance of subsequent first-order models trained on their alignments [19]. We use one cycle learning rate schedule (OCLR) with a peak LR of around (BLSTM: 4e-4, Conformer: 8e-4) over 90% of the training epochs, followed by a linear decrease to 1e-6 [31, 19]. The sequences are chunked into (Conformer: 400, BLSTM: 64) frames with a shift of (Conformer: 200, BLSTM: 64). An Adam optimizer with

Table 1: Comparison for LBS 960h between discriminative models with CTC and HMM topologies, using various label posterior and transition scales, α and β , respectively. In addition to the WER using a 4-gram LM, we also show the time stamp error (TSE) of their alignment with respect to a GMM alignment, the percentage of silence (Si) in HMM and blank (B) in CTC, as well as the average phoneme duration(Phon).

Model	α	β	Align model on train 960h			WER [%]	
			TSE [ms]	Si/B [%]	Phon.[ms]	dev-other	test-other
GMM	1.0	1.0	0	17.5	85.0	19.8	-
P-HMM	0.1	0.1	311	60.3	40.0	12.5	12.2
			189	53.7	47.5	9.5	10.0
			114	3.4	100.4	9.2	9.1
			48	15.1	88.2	8.3	8.8
			77	48.1	53.9	8.8	9.3
	0.7	1.0	85	46.5	55.0	8.4	8.6
			47	27.1	75.8	8.5	8.9
			101	55.5	45.9	9.2	9.8
			176	58.5	43.0	15.0	15.8
			176	58.5	43.0	15.0	15.8
CTC	-	345	0.2	103.0	26.0	26.3	
		240	0.5	102.2	13.4	13.5	
		180	1.6	100.9	11.2	11.6	
		116	6.1	96.6	9.5	9.7	
		38	54.2	47.5	8.0	8.7	

Table 2: Similar experiments as explained in Table 1 for a subset of models using both dev and test sets for SWB 300h.

Model	α	β	Align model on train 300h			WER [%]	
			TSE [ms]	Si/B [%]	Phon.[ms]	hub5*00	hub5*01
GMM	1.0	1.0	0	25.1	86.5	18.9	-
P-HMM	1.0	0.1	79	50.4	57.2	14.9	14.1
			62	27.2	84.0	13.7	13.0
	0.7	1.0	71	20.5	91.7	14.0	13.4
			117	17.4	95.3	14.7	14.3
CTC	-	1.0	58	54.9	51.9	13.7	13.0
		0.7	129	19.2	93.3	14.2	13.7
		0.5	240	12.3	101.1	16.7	16.6

Nesterov momentum, together with optimizer epsilon of 1e-8 are used [32]. The Conformer and BLSTM models are trained for (SWB: 50, LBS: 15) and (SWB: 50, LBS: 20) epochs, respectively. We use a single consumer GPU for 5 to 15 days depending on the setup. The reference alignments used for calculation of TSE are a triphone GMM [33] for LBS and a tandem GMM [34] for SWB, both with speaker adaptation. Decoding follows a time-synchronous beam decoding with dynamic programming principle using lexical prefix tree. All decoding experiments use the official 4-gram language model provided with the respective tasks. For the real time factor measurement experiment, we used an AMD CPU (released 2021), with 2 logical cores. For further details on training hyper parameters and decoding settings, we refer to an example of our configuration setups¹.

4. Evaluation

4.1. Alignment Quality

Our alignment models of Sec. 2.4.1 are trained from scratch with the full-sum training criterion of Eq. (5). A CUDA implementation of the dynamic programming formulation of the forward-backward algorithm computes the state marginals used in the calculation of the loss [35]. The kernel is agnostic to the label topology, enabling a close comparison between the two models.

The P-HMM formulated in Eq. (4) consists of a label posterior and a transition probability. By scaling each component, one can control their distribution shape and therefore their

¹ <https://github.com/rwth-i6/returnn-experiments>

Table 3: ASR performance of factored hybrid HMM (FH) and mRNN-T in terms of WER for LBS960 and SWB300h, on the alignments presented in Tables 1 and 2.

Model	Align Model on LBS 960				WER [%]	
	α	β	TSE [ms]	Si/B [%]	dev-other	test-other
FH	1.0	-	77	48.1	7.0	7.6
	0.7	0.1	48	15.1	6.4	6.9
	0.5	-	114	3.4	6.8	7.2
mRNN-T	1.0	-	40	61.3	6.8	7.1
	0.7	-	610	11.4	7.5	7.9
	0.5	-	609	17.8	20.0	22.3

Table 4: Similar experiments explained in Table 3 for SWB300.

Model	Align Model on SWB 300h				WER [%]	
	α	β	TSE [ms]	Si/B [%]	hub5'00	hub5'01
FH	1.0	-	79	50.4	11.7	11.6
	0.7	0.1	62	27.2	11.4	11.2
	0.5	-	117	17.4	11.9	11.5
mRNN-T	1.0	-	62	65.3	12.5	12.2
	0.7	-	127	65.2	13.4	12.3
	0.5	-	236	65.3	14.4	13.7

frame-level log-linear contribution. This has an effect on the certainty with which the model would choose to stay in a label segment or emit a new label. Therefore, there is a close relation between the input frame shift (classic 10ms against our 40ms), and the choice of aforementioned scales. Viewing the training as a general expectation-maximization (EM) procedure, the scales are then applied before passing the scores to the alignment FSA, i.e. at the expectation step, guaranteeing the local normalization constraint at the maximization step.

Differently to the WER that is a well-defined evaluation metric, it is not clear how to evaluate the quality of the alignment of our models. Due to the lack of a proper ground-truth, a set of different measurements in our case is considered. In addition to the WER of the alignment model, we consider the time stamp error (TSE), i.e. the mean absolute distance (in milliseconds) of word start and end positions against a reference GMM alignment, irrespective of the silence [16, 36]. Furthermore, we include the average phoneme duration, as well as the percentage of silence and blank frames for HMM and CTC alignments, respectively. We then choose within each topology the model with best combination of TSE and WER, and train a first-order model using its alignment. As a final measure, we consider the WER of this system, reported in Sec. 4.2. There are different aspects to be taken into account when comparing the statistics and measures between the two models. Due to the ambiguous role of the blank label in CTC topology, the average phoneme duration is not comparable between the two models, since part of the phoneme duration might be consumed by blank in CTC. This applies also to the comparison between silence and blank percentage. Furthermore, when transferring the CTC alignment for RNN-T training, label loops are removed, following the topology. For the statistics reported in Tables 1 and 2, we use the alignment without the mentioned post-processing, and we will report the TSE of the post-processed alignment in Sec. 4.2. The results show that differently to CTC and its peaky alignment, the P-HMM with lowest WER has a phoneme duration and silence percentage that is near to the reference GMM alignment. In the CTC alignment, we observed that the word-starts tend to be shifted forward while the word-ends tend to be shifted backward in comparison to the reference. This may indicate that the label peaks are typically placed around the center of the actual label duration.

We examined different combinations of scales for label posterior and transition probability in HMM, along with the label

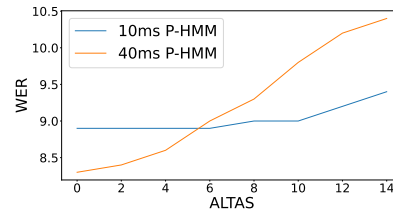


Figure 1: comparison of WER on dev-other between two P-HMMs with 10ms and 40ms frame shift when using acoustic lookahead temporal approximation scale (ALTAS). Note that the scale value represents a different temporal range for each curve.

Table 5: Optimal real time factors (RTF), with distinction between the encoder (Enc.) and search for best 40ms factored hybrid (FH) and mRNN-T models, as well as 10ms FH and traditional hybrid-HMM. Reporting the average number of active state (#S) and lexical prefix-trees (#L) hypotheses at each step, together with WERs and the min and max values for the confidence interval.

Model	Frame Shift [ms]	#S	#L	RTF			WER [%] dev-other	Conf. Int.	
				Search	Enc.	Σ		min	max
Hybrid-HMM	10	120	14	0.02	0.15	0.17	6.6	5.2	8.4
FH	40	410	56	0.04	0.21	0.25	6.5	5.2	8.2
		47	71	0.01	0.11	0.11	6.4	5.2	8.0
mRNN-T	40	11	4	0.05	0.10	0.15	6.8	5.5	8.4

posterior in CTC. The results indicated that the optimal scales of each topology were task-independent, i.e. the same for LBS and SWB. Unsurprisingly, CTC necessitates no scaling of the label posterior, i.e. the optimal value is 1.0. For HMM, the optimal label posterior and transition scales were found to be 0.7 and 0.1, respectively. The optimal label posterior scale for a P-HMM trained with 10ms shift, was shown to be 0.3 [16]. This confirms the relation between the scale and the frame shift. To further verify whether the encoder output with 40ms has less correlated input information at each step, we also use the acoustic lookahead temporal approximation scale (ALTAS) during decoding [37]. In this method the acoustic lookahead score for time $t + 1$ is approximated by the scaled score at time t . This pruning method works well, if the scores between adjacent frames are highly correlated. Therefore, increasing the ALTAS value without degrading WER suggests a stronger correlation among adjacent frames. The faster degradation of the 40ms P-HMM can be seen in Fig. 1.

4.2. ASR Performance

In Tables 3 and 4, it is possible to verify the performance of the models, using the best alignment for each topology from Sec. 4.1. Even though the CTC alignments had slightly better WER and TSE, here the subsequent training favors the HMM approach on both tasks. Finally, we compare the decoding real time factor using a simulated production setup with 75% load on the machine. It is possible to see that the 40ms shift brings up to 50% relative speedup to the FH, surpassing also the classic hybrid-HMM.

5. Conclusions

In this work, we compared two context-independent alignment models with CTC and HMM topologies, and two first-order label context models with monotonic RNN-T and factored HMM topologies. We examined the alignment quality by using various measurements and compared the ASR performance and the real time factor of our best systems during decoding. Moreover, we showed that also the HMM topology can benefit from larger frame shift in the neural encoder architectures, respecting its inherent output independence assumption.

6. Acknowledgements

This work was partially supported by NeuroSys, which as part of the initiative "Clusters4Future" is funded by the Federal Ministry of Education and Research BMBF (03ZU1106DA). Sincere appreciation is extended to Peter Bell (University of Edinburgh) and Lukáš Burget (Brno University of Technology) for insightful conversations and invaluable comments. Authors thank Noureldin Bayoumi and Moritz Gunz for training the CTC and the CART hybrid models, respectively.

7. References

- [1] R. K. Moore and R. Marxer, "Progress and prospects for spoken language technology: Results from five sexennial surveys," *Proc. Interspeech*, pp. 401–405, 2023.
- [2] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [3] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 25–47, 2002.
- [4] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
- [5] J. Lafferty, A. McCallum, F. Pereira *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, vol. 1, no. 2, 2001, p. 3.
- [6] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proc. Interspeech*, 2005, pp. 1117–1120.
- [7] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free MMI," in *Proc. Interspeech*, 2018, pp. 12–16.
- [8] T. Raissi, C. Lüscher, M. Gunz, R. Schlüter, and H. Ney, "Competitive and resource efficient factored hybrid HMM systems are simpler than you think," *Proc. Interspeech*, 2023.
- [9] D. Gillick, L. Gillick, and S. Wegmann, "Don't multiply lightly: Quantifying problems with the acoustic model assumptions in speech recognition," in *Proc. IEEE ASRU*, 2011, pp. 71–76.
- [10] A. Rouhe, T. Grósz, and M. Kurimo, "Principled comparisons for end-to-end speech recognition: Attention vs hybrid at the 1000-hour scale," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 623–638, 2023.
- [11] D. Gimeno-Gómez and C.-D. Martínez-Hinarejos, "Comparison of conventional hybrid and CTC/attention decoders for continuous visual speech recognition," *arXiv:2402.13004*, 2024.
- [12] T. Bayes, "An essay towards solving a problem in the doctrine of chances. by the late rev. Mr. Bayes, FRS communicated by Mr. price, in a letter to John canton, AMFR S;," *Philosophical Transactions of The Royal Society of London*, no. 53, pp. 370–418, 1763.
- [13] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.
- [14] W. Zhou, S. Berger, R. Schlüter, and H. Ney, "Phoneme based neural transducer for large vocabulary speech recognition," in *Proc. IEEE ICASSP*, Jun. 2021, pp. 5644–5648.
- [15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.
- [16] T. Raissi, W. Zhou, S. Berger, R. Schlüter, and H. Ney, "HMM vs. CTC for Automatic Speech Recognition: Comparison Based on Full-Sum Training from Scratch," in *Proc. IEEE SLT*, 2023.
- [17] R. Collobert, C. Puhresch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *arXiv:1609.03193*, 2016.
- [18] A. Tripathi, H. Lu, H. Sak, and H. Soltau, "Monotonic recurrent neural network transducer and decoding strategies," in *Proc. IEEE ASRU*, 2019, pp. 944–948.
- [19] W. Zhou, W. Michel, R. Schlüter, and H. Ney, "Efficient Training of Neural Transducer for Speech Recognition," in *Proc. Interspeech*, Sep. 2022, arXiv:2204.10586.
- [20] T. Raissi, E. Beck, R. Schlüter, and H. Ney, "Context-dependent acoustic modeling without explicit phone clustering," in *Proc. IEEE ICASSP*, 2020.
- [21] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone Speech Corpus for Research and Development," in *Proc. IEEE ICASSP*, 1992.
- [22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," in *Proc. IEEE ICASSP*, 2015.
- [23] P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, and H. Ney, "RETURNN: The RWTH extensible training framework for universal recurrent neural networks," in *Proc. IEEE ICASSP*, 2017, pp. 5345–5349.
- [24] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, "RASR-the RWTH Aachen university open source speech recognition toolkit," in *Proc. IEEE automatic speech recognition and understanding workshop*, 2011.
- [25] W. Zhou, E. Beck, S. Berger, R. Schlüter, and H. Ney, "RASR2: The RWTH ASR toolkit for generic sequence-to-sequence speech recognition," 2023.
- [26] J.-T. Peter, E. Beck, and H. Ney, "Sisyphus, a workflow manager designed for machine translation and automatic speech recognition," in *Proc. EMNLP*, 2018, pp. 84–89.
- [27] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gamma-tone features and feature combination for large vocabulary speech recognition," in *Proc. IEEE ICASSP*, 2007.
- [28] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "SpecAugment on Large Scale Datasets," in *Proc. IEEE ICASSP*, Brighton, UK, May 2019, pp. 6879–6883.
- [29] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020.
- [30] T. Raissi, E. Beck, R. Schlüter, and H. Ney, "Improving factored hybrid acoustic modeling without state tying," in *Proc. IEEE ICASSP*, 2022.
- [31] L. N. Smith and T. Nicholay, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, 2019.
- [32] T. Dozat, "Incorporating Nesterov Momentum into Adam," in *Proc. ICLR*, 2016.
- [33] C. Lüscher, E. Beck, K. Irie, M. Kitzka, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "RWTH ASR systems for LibriSpeech: Hybrid vs attention-w/o data augmentation," *Proc. Interspeech*, 2019.
- [34] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Speaker adaptive joint training of Gaussian mixture models and bottleneck features," in *Proc. IEEE ASRU*, 2015.
- [35] A. Zeyer, E. Beck, R. Schlüter, and H. Ney, "CTC in the context of generalized full-sum HMM training," in *Proc. Interspeech*, 2017, pp. 944–948.
- [36] X. Zhang, V. Manohar, D. Zhang, F. Zhang, Y. Shi, N. Singhal, J. Chan, F. Peng, Y. Saraf, and M. Seltzer, "On lattice-free boosted MMI training of HMM and CTC-based full-context ASR models," in *Proc. IEEE ASRU*, 2021.
- [37] D. Nolden, R. Schlüter, and H. Ney, "Advanced search space pruning with acoustic look-ahead for WFST based LVCSR," in *Proc. IEEE ICASSP*, 2013.