IMPROVING FACTORED HYBRID HMM ACOUSTIC MODELING WITHOUT STATE TYING

Tina Raissi¹, Eugen Beck², Ralf Schlüter¹,², Hermann Ney¹,²

¹Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Germany
²AppTek GmbH, Aachen, Germany
{raissi,schueter,ney}@cs.rwth-aachen.de, ebeck@apptek.com

ABSTRACT

In this work, we show that a factored hybrid hidden Markov model (FH-HMM) which is defined without any phonetic state-tying outperforms a state-of-the-art hybrid HMM. The factored hybrid HMM provides a link to transducer models in the way it models phonetic (label) context while preserving the strict separation of acoustic and language model of the hybrid HMM approach. Furthermore, we show that the factored hybrid model can be trained from scratch without using phonetic state-tying in any of the training steps. Our modeling approach enables triphone context while avoiding phonetic state-tying by a decomposition into locally normalized factored posteriors for monophones/HMM states in phoneme context. Experimental results are provided for Switchboard 300h and LibriSpeech. On the former task we also show that by avoiding the phonetic state-tying step, the factored hybrid can take better advantage of regularization techniques during training, compared to the standard hybrid HMM with phonetic state-tying based on classification and regression trees (CART).

Index Terms— CART-free hybrid HMM, BLSTM acoustic model, regularization, Switchboard, LibriSpeech

1. INTRODUCTION

In automatic speech recognition (ASR) with the hybrid neural network/hidden Markov model (NN/HMM) approach [1], the co-articulation effect is addressed by considering tied allophone states in right and left context, known also as senones. The parameter tying relies generally on the state clustering done by classification and regression trees (CART) [2], and requires a frame-wise alignment of the input speech signal to a hidden HMM state sequence. This alignment, can be obtained from a Gaussian Mixture Model (GMM) system, and is used for both estimation of CART and frame-wise cross-entropy (CE) training of the neural network. The NN component is used for the substitution of the likelihood of a given feature vector by the normalized scaled posterior probability over the tied state label. The overall system has separate components and requires different sources of knowledge. Moreover, having tied state labels as the NN target can lead to overfitting problems due to the arbitrary choices done for the estimation of the clustering [3].

In order to simplify the described pipeline, the recent research focus has shifted to the joint optimization of the different components in a sequence to sequence (seq2seq) framework. In this approach, the speech sequence is directly mapped to the final word sequence, with an optional use of an external LM. In connectionist temporal classification (CTC) [4], the necessity of having an initial frame-wise alignment is relaxed by introducing a blank label and using a sequence-level CE which sums over all possible alignments. The CTC criterion contains an independence assumption at each time frame. The inclusion of the dependency to the full or truncated context is introduced with Recurrent Neural Network Transducers (RNN-T) [5] and Recurrent Neural Aligner [6] with different constraints in how to traverse the alignment lattice. Other prior work focuses on the elimination of state-tying for both hybrid HMM and CTC frameworks using full set of diiphone states [7] or a triphone representation via an embedding network [8]. The aforementioned seq2seq models, are similar to the hybrid approach in providing a statistical model of the posterior distribution over different label topologies and units. This posterior is later used in a (weighted) finite state transducer like structure for decoding. However, they do not use a local normalization and the correct way to combine an external LM is still an open question [9, 10, 11]. The modular formulation in the hybrid approach results in an intrinsic robustness for tasks such as domain adaptation and for low resource ASR tasks [12]. Furthermore, seq2seq models require large amounts of data and many additional regularization techniques in order to reach a reasonable performance. Nevertheless, hybrid models with alternative label units and encoder architectures [13, 14] can still have competitive performance compared to highly optimized seq2seq pipelines [15, 16]. We believe that the first step for a possible simplification of the standard hybrid pipeline is the elimination of the state-tying algorithm. Recently, we presented a factored hybrid HMM model [17]. Given the complete set of triphone states, the factored model learns to predict context-dependent posteriors over phonemes, in a multi-task manner. This is achieved by a decomposition of the joint label identity of the center phoneme state and its left-right phonemes. With this previous work, and its extension [18], we have shown that it is possible to eliminate the
state-tying for the determination of the state inventory, and obtain similar performance to a hybrid CART.

Our main contribution with this work is three-fold: (1) We eliminate the dependency to the clustering algorithm entirely by proposing a from scratch pipeline. With that we obtain slightly better performance compared to the standard hybrid using state-tying, (2) We show that our factored hybrid outperforms the standard approach when using the same initial alignment. And, (3) we show that our modeling approach can obtain larger improvement from regularization techniques, confirming the scalability of our approach to a larger task.

2. MODELING APPROACH

We rely on our previously presented approach [17], with the set \{φ_r, σ_c, φ_t\}, of left phoneme, center phoneme state and right phoneme of the aligned triphone at time frame t. We substitute the emission probability of the feature vector x_t resulting from the decomposition of the acoustic component as follows:

\[
p_t(x|φ_t, σ_c, φ_r) = \frac{p_t(φ_t, σ_c, φ_r | x) \cdot p_t(x)}{p(φ_t, σ_c, φ_r)}
\]

The idea is to carry out the factorization into the left-to-right trigram of Eq. (1a), with additional diphone and monophone formulations, as shown in Eq. (1b) and Eq. (1c), respectively.

\[
p_t(x|φ_t, σ_c, φ_r) \sim \frac{p_t(φ_t | σ_c, x) \cdot p_t(σ_c | φ_t, x) \cdot p_t(φ_t)}{p(φ_t, σ_c, φ_r)} \quad \text{(1a)}
\]

\[
p_t(x|σ_c, φ_t) \sim \frac{p_t(σ_c | φ_t, x)p_t(φ_t | x)}{p(σ_c, φ_t)} \quad \text{(1b)}
\]

\[
p_t(x|σ_c) \sim \frac{p_t(σ_c | x)}{p(σ_c)} \quad \text{(1c)}
\]

In practice, the decision rule in the maximum a-posteriori framework, with input feature x_t, and word sequence w_t, is done via log-linear combination of different factors:

\[
\arg \max_{w_t} \left\{ \max_{\{φ_t, σ_c, φ_r\}^T: w_t^N} \left[ \sum_{t=1}^{T} \log p_t(φ_t | σ_c, φ_r, x) - \gamma_r \cdot \log p(φ_r | σ_c, φ_t) + \log p_t(σ_c | φ_t, x) - \gamma_c \cdot \log p(σ_c | φ_t) + \log p_t(φ_t | x) - \gamma_φ \cdot \log p(φ_t) + \beta \cdot T (t, t-1) \right] + \alpha \cdot \sum_{n=1}^{N} \log p(w_n | w_{n-1}^{N-1}) \right\}
\]

Our general formulation is shown in Eq. (2) using prior, transition penalty, and language model (LM) scales γ, β, and α, respectively. Here, we consider a 4-gram LM. The term T is a constant representing the state transition probabilities, with distinct values in case of speech or silence.

3. TRAINING AND DECODING

The common starting point for all experiments is the initialization of the parameters of a context-independent single Gaussian HMM based on a linear segmentation of the input signal. Using the resulting alignment, a GMM monophone system is bootstrapped. We denote the resulting alignment as GMM-Mono. For the CART system, the set of context-dependent tied states are first clustered in order to build a GMM triphone system. After the inclusion of the speaker adaptation, another CART for clustering the final HMM state inventory is estimated. Each described step consists of different iterations of parameter re-estimation and subsequent re-alignments. We denote the intermediate alignment obtained from the speaker-adapted triphone system as GMM-Tri. For the final best alignment, which we denote as Tandem, first a tandem model with speaker adaptation is trained, followed by minimum phone error sequence discriminative training [19]. All the described steps above are necessary for the CART based system in order to reach a good performance. For each of the alignments described above, we compare Word Error Rates (WERs) of our factored hybrid models. We use GMM-Mono and Tandem alignments for Switchboard 300h, and GMM-{Mono, Triphone} alignments for Librispeech 960h.

The triphone factored hybrid, as described in Eq. (1a), has three separate softmax layers: two of them have as target the left and right phonemes and the third is over the center phoneme states. The context-dependency for the right phoneme and center state is fed via a soft embedding. The overall training procedure consists of multiple stages. In each stage, we increase the model complexity as follows. We start with a monophone model without feeding any context. In the following stages, we increase the context-dependency by feeding the context via phoneme and phoneme state embeddings [17]. The effect of the multi-stage training is reported in Table 1.

In addition to the common regularization techniques described later in [3], we focused on the effect of the following two techniques. **Label Smoothing (LS)** is usually used in order to contrast the model’s overconfidence about the aligned target [20]. The minimization of the cross-entropy in LS is done with respect to a soft target. The target is obtained by interpolating a uniform distribution over other classes with the aligned target. Differently to the monophone targets, where the problem is well-defined, the application of LS is less straightforward when context-dependent targets are considered. Since the posterior over the untied triphone state corresponds to the combination of the factored posteriors, the idea is to keep a hard target for the center phoneme state and apply LS only on the context. Consequently, starting from the diphone stage we turn off LS on the center state output.
Experimental results, as reported in Table 2 for a CART-free pipeline starting with the GMM-Mono alignment confirm the following: (1) The relative improvement obtained from LS in monophone model is larger than in di- and triphone models, (2) continuing with LS on all outputs leads to degradation for diphone model. For CART based models application of LS did not help. In this case, each label corresponds to a diphone model. For CAR T based models application of LS did not help. In this case, each label corresponds to a diphone model. For CAR T based and different factored hybrid (FH) models, evaluated on Hub5’00 with 4-gram LM. Rows with the same color correspond to one multi-stage pipeline, meaning that n-phone model is initialized with (n-1)-phone model.

Table 2: Label Smoothing in SWB 300h for targets in the output layers of CART based and different factored hybrid (FH) models, evaluated on Hub5’00 with 4-gram LM. Rows with the same color correspond to one multi-stage pipeline, meaning that n-phone model is initialized with (n-1)-phone model.

## Table 1: Comparison of the performance of factored triphone models trained on SWB 300h initialized with random parameters and initialized with a diphone model, with no label smoothing, no data augmentation and no realignment. Evaluation is on Hub5’00 with 4-gram LM.

| Model | Alignment | Pretrained | WER[%] |
|-------|-----------|------------|--------|
| CART  | Tandem    | no         | 13.4   |
| FH    | GMM-Mono  | yes        | 13.7   |
| Triphone | yes      | 13.5   |
|       | no        | 15.2   |
|       | yes       | 14.5   |

* Also initialized with monophone using label smoothing on all targets.

CART system for Librispeech follows the system described in [27]. The SWB systems for both CART and factored hybrid share the same setting and similar model sizes. We kept the same number of parameters for the BLSTM encoder for all experiments, comprising 6 forward and backward layers of size 512 with 10% dropout probability [28]. The input speech signal to the encoder is represented by (SWB: 40 and LSB: 50) dimensional Gammatone Filterbank features [29], extracted from 25 milliseconds (ms) analysis frames with 10ms shift. Concerning the factored architectures, the one-hot encoding of the left and right phonemes, and the center phoneme states are projected by using linear layers of dimension 64 and 256, respectively. A two-layer perceptron in all factored models are used to combine the embedding and encoder outputs. The state inventory of both corpora consists of the complete set of triphone states, corresponding to a tripartite 0-1-2 HMM label topology for each phoneme in context, and a special context-independent silence state. Differently to the standard LS phoneme inventory, we removed stress markers from all phonemes. For the standard hybrid, a set of (SWB: 9001, LSB: 12001) CART labels are considered. All models are trained with frame-wise CE using external alignments presented earlier in Sec. [3]. In order to speed-up the training, the sequences are divided into chunks of length 128 with 50% overlap, and fed to the network in batches of 10k. Our models on all tasks are trained using an Adam optimizer with Nesterov momentum [30]. We also use Newbob learning rate (LR) scheduling with an initial LR of $1e^{-3}$, and a decay factor of 0.9 based on the frame error rates. The minimum LR value is set to (SWB: $2e^{-5}$, LSB: $1e^{-5}$). For the factored hybrid the reference error value results from the summation of the frame errors from the three outputs. In addition to label smoothing with scale factor 0.2, we use also other regularization techniques: $L_2$ weight decay with a scale of 0.01, gradient noise [31] with a variance of 0.1, and focal loss factor of 2.0 [32]. CART model converges after (SWB: 24, LSB: 6.25) epochs. Factored models are generally trained up to maximum 30 epochs for each stage. The optimal tradeoff between number of epochs in combination with multi-stage phonetic training has still not been explored and is left for future work. For recognition we use 4-gram [33] and LSTM language models [34]. We also include a second pass rescoring with a Transformer (Trafo) LM for one of our experiments [35]. In addition to the LM scale, the separate transition penalty and prior scales of Eq. (2) are tuned. The explicit context modeling and the decomposition into context-dependent sub-architectures of the factored model require a careful implementation for the computation of the acoustic scores. The search space organization in case of both models follows the lexical prefix tree structure, where by means of dynamic programming, a time-synchronous beam decoding is carried out [36, 37]. At each time frame a set of path hypotheses are propagated and pruned to the most promising ones. Thus, not all the left-center context pairs are

---

1 See code: https://github.com/rwh-iso/returnn-experiments/tree/master/2022-swb-factored-hybrid.
Table 3: The effect of Specaugment (Specaug), label smoothing (LS) and realignment (Realign) in SWB 300h for factored hybrid models compared to the standard C/ART based model, evaluated on Hub5’00 with 4-gram LM.

| Specaug | LS | Model | Alignment     | Realign | [% WER] |
|---------|----|-------|---------------|---------|---------|
| no      | no | CART  | Tandem        | no      | 17.3    |
|         |    |       |               |         | 15.0    |
|         |    |       |               |         | 13.4    |
| no      | yes| FH    | Tandem        | no      | 15.1    |
|         |    |       |               |         | 14.0    |
|         |    |       |               |         | 13.5    |
| yes     | yes| CART  | Tandem        | no      | 15.0    |
|         |    |       |               |         | 14.0    |
|         |    |       |               |         | 13.4    |
| yes     | yes| FH    | Tandem        | no      | 14.6    |
|         |    |       |               |         | 13.6    |
|         |    |       |               |         | 13.1    |
|         |    |       |               | yes     | 16.1    |
|         |    |       |               |         | 14.3    |
|         |    |       |               |         | 13.8    |

4. EXPERIMENTAL RESULTS

The effect of label smoothing and Specaugment on both standard and factored hybrid are compared in Table 3. It is possible to see that the standard hybrid improves only by 2%, whereas the factored hybrid trained with the same alignment benefits up to 5% relative under similar conditions, and 7% relative with one realignment step. The factored model using GMM-Mono alignment has similar performance to standard C/ART based hybrid, making C/ART an optional feature for hybrid models. The results in Table 3 show that higher n-gram context and realignment consistently give improvement in all experiments. In Table 4 we present our baseline for the LibriSpeech 960h task. The experiments did not use the two mentioned regularization methods. However, we can see similar behavior to our SWB baseline, where in absence of label smoothing and Specaugment the quality of the initial alignment decides for the final performance.

Even though the focus of our work is on the elimination of state-tying, we did not limit ourselves to mere comparison of our approach to the standard hybrid. We also trained an enhanced model using speaker adaptation with i-vectors [38], and in Table 5 we compare our Word Error Rates using both 4-gram and LSTM LMs with other state-of-the-art results in the literature. It is possible to see that our approach outperforms the hybrid model that is in addition using sequence discriminative training (sMBR) [38] on averaged HUB5’00, by 1.9% relative. The flat-start hybrid model using no state-tying and trained with lattice-free maximum mutual information (LF-

Table 4: Our baseline results with Librispeech 960h and evaluated on both dev and test sets.

| Model  | Alignment | %WER |
|--------|-----------|------|
| CART   | GMM-Mono  | 4.1  |
| FH     | GMM-Mono  | 3.8  |

Table 5: Evaluation of SWB300h systems on different subsets of HUB5’00 and HUB5’01 from the literature.

| Work | Model          | #Eps | LM  | HUBS’00 | HUBS’01 |
|------|----------------|------|-----|---------|---------|
| FH   | AED            | 750  | RNN | 6.8     | 14.1    |
| FH   | Hybrid (LF-MMI) | 4   |     | 9.3     | 18.9    |
| FH   | Hybrid+ivec+sMBR | - | LSTM | 6.7     | 14.7    |
| FH   | RNN-T          | 100  |     | 6.3     | 13.1    |
| FH   | Phon-Trans     | 7.6  | 15.4 | 8.4     | 10.6    |

5. CONCLUSIONS

In this paper, we have shown that the elimination of explicit state-tying from hybrid HMM models results in performance gains across multiple tasks. Our factored hybrid approach works better in combination with popular regularization techniques compared to a hybrid system using state-tying. This work also demonstrates that the hybrid approach is still competitive with nowadays popular sequence to sequence models.

6. ACKNOWLEDGMENTS

This work was partly funded by the Google Faculty Research Award for “Label Context Modeling in Automatic Speech Recognition”. This work was partially supported by the project HYKIST funded by the German Federal Ministry of Health on the basis of the decision of the German Federal Parliament (Bundestag) under funding ID ZMVII-2520DAT04A. We would like to thank Christoph M. Lüscher and Alexander Gerstenberger for providing us the LibriSpeech CART baseline and performing the lattice rescoring experiment with Transformer, respectively.
