Acoustic Look-Ahead for More Efficient Decoding in LVCSR

D. Nolden, R. Schlüter, H. Ney

Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{nolden, schlueter, ney}@cs.rwth-aachen.de

Abstract

In this paper we propose novel approximations of a generalized acoustic look-ahead to speed up the search process in large vocabulary continuous speech recognition (LVCSR). Unlike earlier methods, we do not employ any phoneme- or syllable level heuristics. First we define and analyze the *perfect* acoustic look-ahead as a simple pre-evaluation of the original acoustic models into the future. This method is very slow, but reveals the best possible impact on the search space that can be achieved through acoustic look-ahead. In a second step, we derive efficient and simple *approximative* look-ahead models from the perfect models. We show that the approximative models compare well to the perfect models regarding the search space, and that the approximative models significantly improve the efficiency in comparison to the baseline, without any negative effect on the precision.

Index Terms: speech recognition, search, acoustic look-ahead, efficiency

1. Introduction

Most state of the art LVCSR decoders follow the time synchronous beam-search approach, based on a hidden Markov model (HMM) acoustic model (AM), in combination with an n-gram language model (LM).

In dynamic network decoders, the LM and AM are combined dynamically. The acoustic model is used to build a compact HMM search network representing all the words in the vocabulary, and the LM dependencies are maintained by appropriate dynamic management of state hypotheses [1].

Static decoders on the other hand, usually based on the weighted finite state transducer (WFST) approach [2], combine the AM and LM statically by building one huge HMM search network representing both models.

In all state of the art approaches, the future LM probabilites are integrated early into the search process to focus the search towards the most promising branches of the network. In dynamic decoders, the future LM probabilities can be integrated efficiently by building LM look-ahead structures on-demand [3]. In WFST decoders, LM look-ahead is performed implicitly, by pushing the weights within the automaton toward the root.

Orthogonally to the LM look-ahead, the search can also be focused onto promising branches of the search network by using the knowledge of future *acoustic* observations, forming an *acoustic look-ahead*.

Earlier, acoustic look-ahead has been realized using the phoneme look-ahead technique, which uses the knowledge about phonemes in the search network to prevent transitions into phonemes that are unlikely given a specific range of future acoustic observations [4] [5]. However, phoneme look-ahead is not compatible with a fully compressed search network based on allophones with tied sub-states, because in such a network, clear transitions between individual phonemes do not exist. Furthermore, since classical phoneme look-ahead is only applied to transitions between phonemes, only a small fraction of all transitions are actually affected. The phoneme look-ahead is applied only during a separate pruning step, so its interaction with the LM look-ahead is not optimal. During each pruning step, either LM- or phoneme look-ahead is considered independently.

In [6] acoustic look-ahead is performed by building a syllable lattice in a first pass, and using that lattice in a second pass to assign look-ahead scores to all state hypotheses during search. The downsides of this approach are that an additional pass is required to compute the syllable lattice, and that heuristical knowledge about the syllables is not available for most languages without additional effort.

In this paper, we propose efficient acoustic look-ahead techniques that cleanly integrate into the standard search process equivalently to LM look-ahead, with optimal interaction between between both look-ahead knowledge sources, and without an additional pruning step. The acoustic look-ahead is applied onto all state hypotheses during acoustic pruning, it does not require any heuristical knowledge about syllables or phonemes, it works online in a single pass, and it is compatible with a fully compressed search network. The look-ahead models are completely derived from the original acoustic models, and the techniques are relevant for both dynamic and static decoders.

In the following, we first review the structure of a dynamic decoder and the application of the LM look-ahead during search. Then we define the *perfect* acoustic look-ahead as a simple pre-evaluation of the original acoustic models into the future, and in the following we derive more efficient approximative models. We evaluate all acoustic look-ahead methods regarding the average number of active state hypotheses during decoding (search space), real time factor (RTF) and word error rate (WER) on a LVCSR task consisting of 192 minutes of speech.

2. Review of Decoding and Pruning

Although acoustic look-ahead is generally relevant for any kind of decoder, we will base our experiments on a dynamic network decoder where LM and AM are combined dynamically.

The goal of the decoder is, for a given sequence of T acoustic observation vectors $x_1^T = x_1...x_T$, finding the most probably spoken word sequence w_1^N of N words, according to the underlying AM and LM:

$$[w_1^N]_{opt} = \underset{N,w_1^N}{\text{argmax}}\{p(w_1^N) \cdot p(x_1^T|w_1^N)\} \tag{1}$$

Where the LM models the probability of the word sequence $p(w_1^N)$, and the AM models the probability of the acoustic observations given the word sequence $p(x_1^T|w_1^N)$.

For LVCSR, the modeling of the acoustic observations has to be broken up into smaller sub-units of words, usually context dependent phonemes (allophones). Each allophone is modeled by a sequence of HMM states, whereas the emission probability of each HMM state is modeled by one of a limited number of emission models.

In dynamic decoders, usually a static HMM search network, representing all the words in the vocabulary and compressed according to the emission models, is expanded.

The active search space is managed dynamically, by propagating state hypotheses through the search network and handling ending words and path recombination according to the

Viterbi approximation. Each state hypothesis (s, h) stands for a path ending in state s with the predecessor word history $h = u_1^N$.

Due to pruning, only a small fraction of the possible state hypotheses is active at each timeframe. $Q_h(t, s)$ denotes the probability of the best path through the HMM network that ends at timeframe t in state s with LM context h, and is defined for all active state hypotheses (s, h):

$$Q_h(t,s) = \max_{\substack{s_1,\ldots,s_t \\ :s_1=0,s_t=s}} \underbrace{\prod_{f=1}^{t} \{\underbrace{p(x_f|s_f)}_{\text{emission}} \cdot \underbrace{p(s_f|s_{f-1})}_{\text{transition}}\}}_{\text{AM}} \cdot \underbrace{p(h)}_{\text{LM}} \tag{2}$$

Where the emission- and transition probabilites are modeled by the AM, and the probability of the entire predecessor word sequence $p(h)$ is modeled by the LM. Only state sequences s_1^t that are consistent with the HMM of the predecessor word sequence h, that start in the root state $s = 0$, and that end in the corresponding state s are considered. h denotes the *complete* predecessor word sequence, not a truncated sequence as common in similar notations.

Acoustic pruning is used to reduce the number of active state hypotheses by applying a beam around the best state hypothesis. Let $Q_{max}(t)$ be the probability of the overall best state hypothesis of timeframe t:

$$Q_{max}(t) = \max_{s,h} Q_h(t,s) \tag{3}$$

During acoustic pruning, only those state hypotheses (s, h) are preserved that have a probability higher than $Q_{max}(t) \cdot f_{AC}$:

$$Q_h(t,s) > Q_{max}(t) \cdot f_{AC} \tag{4}$$

Where $f_{AC} < 1$ is the acoustic pruning threshold.

Acoustic pruning is performed at each timeframe after the emission probabilities $p(x_t|s)$ of the last state on the hypothesized path (see Equation 2) were computed.

Additionally, *histogram pruning* is used to limit the total number of active state hypotheses, using a dynamic acoustic pruning threshold that is automatically tuned to a value so that the number of active state hypotheses stays below a specified maximum.

When reliable look-ahead information is available, it makes sense to add an additional acoustic pruning step *before* computing the emission probabilities, but after having computed the look-ahead probabilities, because the computation of emission probabilities is one of the major bottlenecks during decoding.

Therefore we add an *early acoustic pruning* step, which works exactly as the standard acoustic pruning, but is performed *before* emission probabilities are computed.

3. LM Look-Ahead

With LM look-ahead, the probabilities of reachable word ends are included during the acoustic pruning. Let $W(s)$ be the set of word ends that are reachable from state s. The LM look-ahead probability is the probability of the best word that can be reached from state s for history h:

$$\pi_h(s) := \max_{w \in W(s)} p(w|h) \tag{5}$$

The state hypothesis probability extended by the LM look-ahead probability is:

$$\tilde{Q}_h(t,s) = Q_h(t,s) \cdot \pi_h(s) \tag{6}$$

And the acoustic pruning, as per Equation (3) and (4), is then performed on $\tilde{Q}_h(t,s)$ instead of $Q_h(t,s)$. The full look-ahead probabilities can be computed efficiently [3].

Through LM look-ahead, the beam search is focussed onto those branches of the network which are most promising according to the LM. As an effect, the size of the active search space is significantly reduced at equal acoustic pruning constraints, and a tighter acoustic pruning becomes feasible without additional errors [3].

4. Perfect Acoustic Look-Ahead

As seen in Equation 2, the probability assigned to each state hypothesis is composed out of one component from the AM, and one component from the LM. When using LM look-ahead, the LM component is extended by the most probable follow-up word reachable from a specific state during pruning (see Equation 6), thus *future* LM information is incorporated to focus the search.

Likewise to LM look-ahead, we can use future information from the AM to extend the AM component during pruning. A *perfect* acoustic look-ahead simply evaluates the AM by a specific number L of timeframes into the future, and uses the most probable successor path to focus the search.

Let $\gamma(t, s)$ be the probability of the most probable HMM search path of length L that starts at timeframe $t + 1$ behind state s:

$$\gamma(t,s) = \max_{s_1,\ldots,s_L:s_0=s} \prod_{f=1}^{L} \{\underbrace{p(x_{t+f}|s_f)}_{\text{emission}} \cdot \underbrace{p(s_f|s_{f-1})}_{\text{transition}}\} \tag{7}$$

We integrate the acoustic look-ahead probability $\gamma(t, s)$ into the acoustic pruning by applying it on top of the LM look-ahead (see Equation 6):

$$\dot{Q}_h(t,s) = \tilde{Q}_h(t,s) \cdot \gamma(t,s)^\alpha \tag{8}$$

Where α is a scaling exponent which is required to keep the balance between LM look-ahead and acoustic look-ahead intact.

The acoustic pruning, as per Equation (3) and (4), is then performed on $\dot{Q}_h(t,s)$ instead of $Q_h(t,s)$.

5. Approximative Acoustic Look-Ahead

The evaluation of state emission models is very expensive, thus a pre-evaluation of the original acoustic models can hardly improve the efficiency. To improve the efficiency, either a *temporal* approximation on the time axis, or a *model* approximation based on the look-ahead models can be applied.

5.1. Temporal Approximation

For a very short look-ahead interval $L = 1$, the locality of hidden Markov models and of acoustic signals can be exploited to approximate the acoustic look-ahead. Since hidden Markov models allow a loop transition, one of the possible successors of a state is always the same state, which makes the states own model an approximation of the look-ahead model. When the HMM employs state repetitions, in many cases the forward-transition will lead to an equal emission model too, and during training, adjacent state models are often aligned with similar acoustic observations, both of which further increases the validity of the own emisson model of a state as a look-ahead model.

HMM locality can be exploited to approximate the acoustic look-ahead of depth $L = 1$ by using the acoustic model of the predecessor state s:

$$\gamma_1(t,s) = p(x_{t+1}|s)^\alpha \tag{9}$$

Such approximated look-ahead is much more efficient than perfect look-ahead with depth 1, because no iteration through multiple successor states is required. However, one additional probability $p(x_{t+1}|s)$ needs to be computed.

Since each feature vector x_i accounts only for a very short duration, there is a certain locality in the acoustic feature vectors: The distance between x_i and x_{i+1} can be expected to be smaller than the distance between two random feature vectors, which makes x_i an approximation of x_{i+1}.

By additionally exploiting the locality of the acoustic feature vectors, we can omit the computation of the additional probability, by simply re-using the probability of timeframe t:

$$\gamma_2(t,s) = p(x_t|s)^\alpha \tag{10}$$

Technically, the probability $p(x_t|s)$ is not a *look-ahead* probability, because it neither incorporates acoustic feature vectors, nor acoustic models from the *future*. Under the assumption of temporal similarity in the HMM and in the feature vectors, this probability still approximates the perfect acoustic look-ahead of depth 1 to a certain degree. $p(x_t|s)$ can be used without any runtime overhead, because the probability is always computed while evaluating the acoustic models (see Equation 2). We have investigated look-ahead based on γ_1 as well as γ_2, and did not observe any differences in precision, while γ_2 is much more efficient, thus in the following we will use the term *temporal approximation* to refer to γ_2.

5.2. Model Approximation

The simplest approximation to the models can be achieved by replacing the state emission models used for look-ahead as per Equation (7) with simpler models that can be evaluated more efficiently.

However, to compute the look-ahead probability with depth 1, an iteration through all sucessor states and an evaluation of all their emission models is still required. Therefore, and to have a better control of the overall number of models, it is desirable to employ specific look-ahead models that combine the acoustic models of all direct successor states into one single simplified model.

Let M be the desired number of simplified models. The mapping $m(s) \in 1, ..., M$ assigns a simplified model to each state s. The simplified models model the probability $p(x|m')$ of observation x given the simplified model m'. The simplified models have to be chosen so that:

$$p(x|m(s)) \approx \max_{s'} p(x|s') \cdot p(s'|s) \qquad (11)$$

Since we want the model evaluation to be as efficient as possible, we choose single-gaussian models instead of mixture-models, and since our models employ a globally pooled covariance matrix, deriving the models is straight-forward: Based on a single-gaussian version of the original emission models, we use an iterative estimation maximization algorithm to repeatedly assign the best matching model $m(s)$ to each state s of the HMM network, and then re-estimate the density of each model m', iteratively minimizing the mismatch defined by Equation 11.

The overall effort is linear in the number of models M, the number of iterations, the size of the HMM network and its branching factor.

The acoustic look-ahead probability of depth 1 based on the simplified models then is:

$$\gamma_3(t, s) = p(x_{t+1}|m(s))^\alpha \qquad (12)$$

5.3. Combination

Since temporal approximation and model approximation each have specific advantages and disadvantages, they can be combined to form an acoustic look-ahead that is more precise than each method individually. The advantage of temporal approximation is that it uses much more complex models, the disadvantages are that it does not really use information from the future and its probability can not be included during *early* acoustic pruning. Model approximation has the advantage that its models are constructed from *future* emission models, are matched with future acoustic observation vectors, and its probabilities can be used during early acoustic pruning, which can reduce the number of real emission probabilities that need to be computed, with the main disadvantage that the approximated models are very simple and agressively tied. Both methods can be combined by multiplying the look-ahead probabilities γ_2 and γ_3.

5.4. Normalized Scale

The optimal acoustic look-ahead scale α is highly dependent on the depth L, because a certain balance needs to be maintained between the LM look-ahead and the acoustic look-ahead. When the overall probabilities assigned through the acoustic look-ahead are much stronger than those from the LM look-ahead, the search is focussed into directions that are promising regarding the AM, but not promising enough regarding the LM. On the other hand, if the probabilities assigned through acoustic look-ahead are too weak, the search stays too unfocused. For example, when fixing $\alpha = 1$, then the optimal observed look-ahead depth in combination with tight acoustic pruning is $L = 7$ (see Section 6). For all depths L, we have observed that the optimal scale is $\alpha = \beta/L$, where β is a constant base scale which defines the depth towards which the look-ahead is normalized, and the ideal scale on our example is $\beta = 7$. Generally, the ideal base-scale β is highly dependant on the quality of the look-ahead models and on the acoustic pruning threshold. Under tight acoustic pruning constraints, the focussing of the search space through acoustic look-ahead seems to have a higher value than under relaxed pruning constraints, where it can even lead to a slightly increased error, probably due to inbalance between acoustic and LM look-ahead.

6. Experimental Results

We perform all experiments on a modified variant of the RWTH Aachen open source speech recognition software. We use a speaker independent recognition system on the EPPS English 2006 evaluation corpus from the TC-STAR project [7], with a 60k word vocabulary, across word modeling, and a 4-gram LM with a perplexity of 129. The LM is integrated early into the search process through sparse 4-gram LM look-ahead [3]. The acoustic model facilitates triphone models tied by a CART tree, each triphone is modeled by 3 individual HMM states, and each HMM state is repeated twice. The acoustic model utilizes 4501 gaussian mixture models consisting of a total of 880244 mixture densities with a globally tied covariance matrix. We can not compare with classical phoneme look-ahead, because the search network is compressed on a HMM state level, where clear transitions between phonemes do not exist. The corpus consists of 192 minutes of speech recorded from plenary sessions of the european parliament with many different speakers. The RTFs are computed on a machine with AMD Opteron 248 processor with 2.2 GHz and 8 GB of memory.

6.1. Look-Ahead Depth

Figure 1 illustrates the WER and RTF reached with perfect acoustic look-ahead at a histogram pruning limit of 500 states, in dependence from the acoustic look-ahead depth L, with a base scale of $\beta = 7$ (which results in the optimal scale α for all depths). Histogram pruning is used to *fix* the number of active HMM state hypotheses, thereby possible differences in the size of the search space are eliminated, and the resulting WERs become comparable. The step from *no look-ahead* to *look-ahead of depth 1* reduces the WER from 26.8% to 16.2%, and further increases of the depth show much smaller effects.

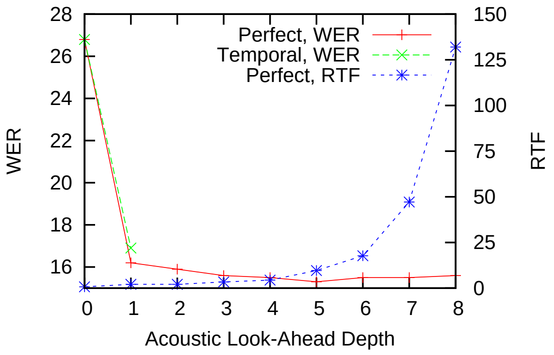

Figure 1: *The WER and RTF at a fixed histogram pruning limit of 500 states, with perfect acoustic look-ahead of varied depth L, and base scale $\beta = 7$.*

The optimal depth is 5, at higher depths the additional depth starts to raise the error rate. Since phonemes have a HMM length of 6 states in our acoustic models, a look-ahead of depth 6 has a higher distance than a full phoneme, which can not be expected to be useful, because the typically most active parts of the search network have such a branching factor that each phoneme is followed by nearly *all* other phonemes. That means that, due to the depth-normalized scale, starting at some depth invaluable distant emission probabilities are added, and at their cost, the more valuable emission probabilities from the *close* future are scaled down. We were able to compensate the negative effect of high depths by reducing the scale α applied to each emission probability by a specific factor with each step, however, we were not able to reach a better WER than at depth 5 in any case.

The temporal look-ahead approximation of depth 1 proposed in Subsection 5.1 achieves nearly the same WER as perfect acoustic look-ahead of depth 1, but without any increase of the RTF in comparison to the baseline.

6.2. Method Comparison

For the perfect acoustic look-ahead, a depth of $L = 3$ was chosen, because the look-ahead is nearly saturated at that depth (see Figure 1), while the computation at higher pruning thresholds is still feasible. For model approximation, a model-count $M = 1000$ was chosen, because this number delivers a good tradeoff between model accuracy and efficiency (each model is evaluated maximally once per timeframe due to caching). For the temporal approximation (see Subsection 5.1) a base scale of $\beta = 2$ was chosen, for model approximation $\beta = 5.8$, and $\beta = 5$ for the perfect acoustic look-ahead with depth 3. For the method combination, we chose an optimal scale of $\beta = 3.5$ for model approximation, and $\beta = 2.5$ for temporal approximation. The scales were chosen experimentally to reach the overall best relationship between WER and RTF.

Figure 2 illustrates the different look-ahead methods and their impact on the WER and the size of the search space. Regarding the relationship between WER and search space, as expected, the perfect acoustic look-ahead is consistently the best method (the slight difference for WER 13.1% can be considered an artifact resulting from the sampling of the pruning thresholds). For the best WER of 13.1%, the combined look-ahead method requires only a search space of 17.2k states, while the baseline requires 30.8k states, which is a reduction of the search space by 44%. For a medium error rate of 13.3% the reduction in comparison to the baseline is 40%. The combined approximative look-ahead method consistently outperforms the baseline as well as both individual approximative methods, and for most error rates reaches about 70% of the search-space reduction achieved by the perfect acoustic look-ahead. The temporal approximation method consistently outperforms the baseline, and performs slightly better than the model approximation method.

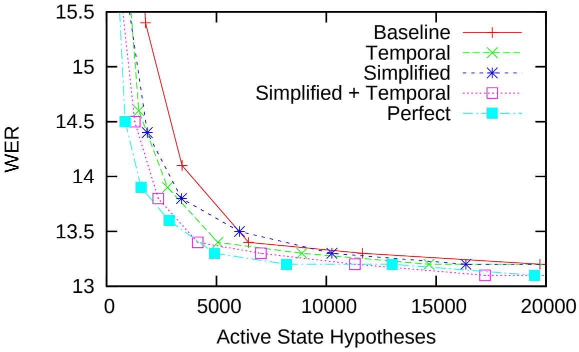

Figure 2: *WER vs. the number of active state hypotheses after pruning, with different acoustic look-ahead methods, under varied acoustic pruning thresholf f_{AC}.*

Figure 3 shows the corresponding relationship between WER and RTF. Perfect acoustic look-ahead can not compete here, because of the many additional evaluations of emission models. The approximative methods consistently outperform the baseline. Unlike the relationship regarding the search space, the model approximation performs better than the temporal approximation, because its probabilities can be considered during early acoustic pruning, which can save expensive evaluations of emission models, and thus has a significant effect on the RTF. The combination of both approximative methods consistently performs best. For the better error rates close to the minimum of 13.1%, a reduction of the RTF by approximately 35% can be achieved, for the higher error rates the reduction is between 50% and 70%.

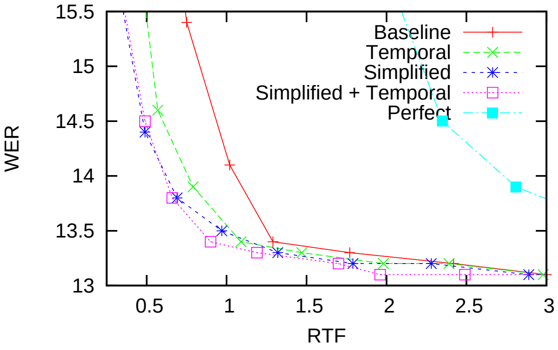

Figure 3: *WER vs. RTF, with different acoustic look-ahead methods, under varied acoustic pruning thresholf f_{AC}.*

7. Conclusions

- Both approximative look-ahead methods combined allow reducing the RTF by 35 to 70% without degrading the precision.
- Even the extremely simple temporally approximated acoustic look-ahead can significantly improve the efficiency without degrading the precision.
- Acoustic look-ahead in general is saturated at a depth of less than 6 timeframes, and most of the positive effect can be achieved already at depth 1.
- There is not much room for further improved approximative acoustic look-ahead methods, because the perfect acoustic look-ahead and its effect on the search space impose a lower bound, and the introduced methods already reach about 70% of the search space reduction achieved through perfect acoustic look-ahead, at negligible costs.

8. Acknowledgements

This work was partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation.

9. References

[1] H. Ney and S. Ortmanns. *Progress in Dynamic Programming Search for LVCSR*. Proceedings of the IEEE, Vol. 88, No. 8, pp. 1224-1240, Aug. 2000.

[2] C. Allauzen, M. Mohri, M. Riley, and B. Roark. *A Generalized Construction of Integrated Speech Recognition Transducers*. ICASSP, 2004.

[3] D. Nolden, H. Ney, R. Schlüter. *Exploiting Sparseness of Backing-Off Language Models for Efficient Look-Ahead in LVCSR*. ICASSP, 2011.

[4] S. Ortmanns and H. Ney. *Look-Ahead Techniques for Fast Beam Search*. Computer Speech & Language, Vol. 14, No. 1, Jan. 2000.

[5] J. A. Sánchez, F. Casacuberta, P. Aibar, D. Llorens, M. J. Castro. *Fast Phoneme Look-Ahead in the ATROS system*. VIII Spanish Symposium on Pattern Recognition and Image Analysis, 1999.

[6] B. Chen, J.-W. Kuo, W.-H. Tsai. *Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription*. ICASSP, 2004.

[7] J. Lööf, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schlüter, and H. Ney. *The 2006 RWTH Parliamentary Speeches Transcription System*. TC-STAR Workshop on Speech-to-Speech Translation, pp. 133-138, Barcelona, Spain, 2006.