COMBINED ACOUSTIC AND LINGUISTIC LOOK-AHEAD FOR ONE-PASS TIME-SYNCHRONOUS DECODING

Xavier L. Aubert *Reinhard Blasig*

Philips Research Laboratories
Weisshausstraße 2, 52066 Aachen, Germany
{Xavier.Aubert, Reinhard.Blasig}@philips.com

ABSTRACT

This paper describes an enhanced pruning technique aimed at a further reduction of the active search space in large vocabulary speech recognition, to speed-up decoding while maintaining the accuracy. The method is based on anticipating both the linguistic and acoustic contribution of a phonetic arc, before expanding that arc in the search. The decoder is based on a time-synchronous beam search and a lexical tree. Cross-word HMMs and M-gram language models are integrated in a single decoding pass. The new algorithm has been evaluated for one-pass trigram decoding of Broadcast news. With respect to the baseline, the search effort can be halved at almost no degradation. When pruning more aggressively to get a speed-up of 10, real-time decoding is achieved on Hub4 evaluation, however, with an increase of the base error rate by one third.

1. INTRODUCTION

This paper presents the integration of an enhanced look-ahead pruning technique in a one-pass time-synchronous decoder and its evaluation on large vocabulary speech recognition tasks. The baseline decoder has been described in [1]. This work aims at reducing the active search space further by paying attention to the acoustic content of the incoming signal, to get a significant speed-up while maintaining the accuracy as high as possible.

The principle of phoneme look-ahead (LA) is not new and has already been applied to earlier versions of similar decoders [2, 3]. It consists in estimating the likelihood of producing a given phone in a short interval ahead of the current time-index, to predict which phonetic arcs are likely extensions of the active partial hypotheses. Unpromising arcs are no longer expanded which reduces the search space to be explored and leads to faster decoding. The phonetic LA scores are usually computed with coarse acoustic models as well as with simplified time-alignments [3, 4] such that the admissibility of this pruning cannot be guaranteed. In practice, a tradeoff has to be found between speed-up and accuracy.

The specific contributions of the present study are (1) the combination of both the linguistic and phonemic LA scores in one single anticipative pruning equation, (2) an improved computation of the LA scores when the phoneme realizations are shorter than the anticipative interval and (3) a systematic evaluation on broadcasts news data.

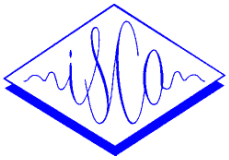
The idea of cutting down the search complexity by means of fast match techniques has been also considered in the framework of other decoding algorithms with distinct solutions. In the IBM stack decoder [5], a fast acoustic match provides a short list of word candidates to extend the most promising theory. Lexical constraints encoded in a prefix tree are used to carry out this fast selection. In the BBN approach [6], the fast match component performs a first decoding pass using some linguistic knowledge as well, the results of which are exploited to prune a second detailed pass based on whole-sentence likelihood estimations. These examples show the wide range of time intervals and knowledge sources that are involved in different applications of the fast match principle.

In the next sections, the baseline decoder is briefly reviewed, making clear that the phonetic arc transitions are the most convenient place to apply a fast acoustic match working in conjunction with the linguistic LA scores. The equations of this combined LA pruning are presented. The phoneme look-ahead component is described and the calculation of the fast phonemic match is explained. The LA scores are no longer time-normalized over the anticipative interval and may involve a subsequent phone to get a better prediction of short phone utterances, i.e. shorter than this interval.

This new pruning has been evaluated in the framework of one-pass 64k trigram decoding. Experimental results are presented for the North American Business (NAB) corpus and the US broadcast news (Hub4) database. In both cases, the efficiency of the decoder is significantly increased, the search space being reduced by a factor of more than two with very little degradation. More aggressive pruning setups have been applied to the Hub4 system for speeding up the first pass by a factor of 10, up to real-time decoding, the base error rate being then increased by about one third, from 21% to 28.3%.

2. BASELINE SEARCH OVERVIEW

The baseline search proceeds from left to right on a strict time-synchronous mode and relies on a prefix tree structure of the lexicon [1]. The search network to be explored is dynamically expanded by re-entering the generic lexical tree whenever a word-end hypothesis is generated and sur-



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

vives pruning. Active hypotheses are made dependent on their word history, the length of which is fixed by the order of the language model (LM). The recombination step that occurs at word ends is very efficiently performed by associating a hash index to each active LM state augmented with the cross-word right phonetic context [1].

Linguistic look-ahead is part of the baseline decoder, a technique that consists in factorizing the word probabilities such that they can be applied incrementally at each phone arc. In the present case, bigram probabilities are spread on demand over the tree using the identity of the immediate predecessor word and this leads to a very effective anticipative pruning that considers the linguistic contribution of an arc before expanding it. The whole search process is controlled by a double beam pruning technique applied on the HMM states and also at word-end nodes, after the exact LM scores have been included. Peaks in terms of large number of active states or word-ends are handled by histogram pruning to confine the search to a maximum number of most promising hypotheses.

Given this architecture, most of the decoding effort is spent in the first phonetic arc generations of the lexical tree and, as shown by our experiments, the whole computational cost is roughly proportional to the average number of expanded arcs, everything being done "on demand". Interacting with the phonetic arc transitions, as it was already done in [2], appears the most convenient choice to integrate a fast acoustic match. However, working at the phone level clearly limits the selection capabilities since only short-range lexical or linguistic constraints can be taken into account and the look-ahead time interval must be kept relatively short, about 10 centi-seconds.

On the other hand, the computation of the acoustic log-likelihood scores often represents an important fraction of the overall decoding cost, especially with continuous mixture HMMs. A number of known methods can be applied to drastically reduce the complexity of mixture calculations [7]. Such techniques are integrated in the current baseline system. Here we focus on cutting down the "pure" search effort by combining the acoustic and linguistic expectations of a lexical tree arc to get more powerful pruning possibilities.

3. COMBINED LINGUISTIC AND ACOUSTIC LA PRUNING

Let "arc" denote any arc of the phonetic tree, "p_arc" its parent arc and $\alpha(arc)$ its phoneme label. Let $W(arc)$ define the set of words that can be reached from this phonetic arc up to the tree leaves.
The overall probability of the best path up to time t and HMM state s depends on the LM history and will be written as $Q_h(t,s)$, where $h = (u,v)$ in case of trigram decoding. The incremental contribution of a successor arc to the cumulated probability of an active path can be decomposed in an acoustic and linguistic part as following.

Given a look-ahead time interval of length Δ_t, the probability estimation that phoneme α is produced by the acoustic vectors $x_{t+1}, ..., x_{t+\Delta_t}$ will be written as $\hat{q}(\alpha; t, \Delta_t)$.
To define the linguistic contribution of an arc, let's first introduce

$$\pi_v(arc) \;:=\; \max_{w \in W(arc)} p(w|v) \;,$$

where $p(w|v)$ is the bigram LM probability conditioned on the predecessor word v. The incremental LM contribution of an arc under bigram is then obtained with

$$\hat{g}_v(arc) = \left(\frac{\pi_v(arc)}{\pi_v(p\_arc)} \right)^{\gamma} ,$$

where γ is the usual language model factor.
Using $\dot{s}$ to denote the ending HMM state s of an arc, the overall probability of the path extended from t to $t+\Delta_t$ with an arc whose phoneme label is $\alpha(arc)$ can be anticipated as

$$\hat{Q}_{u,v}(t+\Delta_t, \dot{s}(arc)) \approx Q_{u,v}(t, \dot{s}(p\_arc)) \cdot \hat{q}(\alpha; t, \Delta_t) \cdot \hat{g}_v(arc)$$

Defining $f_{LA} < 1$ as the look-ahead pruning threshold, the "successor" arc will *not* be further expanded if

$$\hat{Q}_{u,v}(t+\Delta_t, \dot{s}(arc)) < f_{LA} \cdot Q^*(t,\cdot) \cdot \hat{q}^*(\cdot\,; t, \Delta_t) \;,$$

where $Q^*(t,\cdot)$ stands for the current best path score up to t, $\hat{q}^*(\cdot\,; t, \Delta_t)$ is the best phonemic match in the LA interval, the best linguistic LA score being taken as unity. Note that the acoustic LA scores $\hat{q}(\alpha; t, \Delta_t)$ are only used in this beam pruning equation as opposed to the linguistic LA scores that are truly integrated in the scoring process of the main search up to the word-endings.

4. PHONEME LOOK AHEAD

The phoneme LA scores $\hat{q}(\alpha; t, \Delta_t)$ are computed from coarse acoustic HMMs by running a second time-synchronous Viterbi process, offset by Δ_t frames in the future with respect to the "current" time index t.

4.1. Coarse Acoustic HMMs

To keep the additional cost of this LA stage as low as possible, the acoustic models are chosen context-independent with one state per phoneme. These models can be either estimated in a separate training or directly deduced from the detailed context-dependent models by applying clustering on the mixture density components. Using one-state HMMs offers the advantages that (1) the number of density components can be reduced with respect to the standard 3-states topology by eliminating the overlap that exists between distributions of neighboring states and (2) the time-alignment computations are greatly simplified. Similar to [3], our results show that identical LA performances are obtained with one state models made of 50% less densities compared to the 3-states models.

4.2. Computation of Phonemic Match

Basically, the phoneme LA scores are obtained by a straightforward application of Viterbi algorithm on the PDFs of the coarse acoustic HMMs in the time interval $[t+1, t+\Delta_t]$. The length of the look-ahead interval (Δ_t) is

an important parameter, however: if taken very short, the selection capabilities are obviously limited while if taken too long, the LA interval will be longer than the actual phoneme utterances. In principle, the best results would be expected by taking Δ_t equal to the average phone duration, but this value strongly depends, among other factors, on the phoneme category and the speaking style.

Even when using the average phone duration, a problem arises when a phoneme utterance is shorter than this interval. In [3], the authors proceed by normalizing linearly over time the scores of the "short" alignments based on their duration, to cope with the requirements of time-synchronous beam pruning. In this study, another solution has been considered that appears to be more accurate and more robust with respect to the choice of Δ_t. It consists in looking for the best alignment of either one single phone or a "short" realization of that phone followed by any second phone, such that an acoustic match is always consistently computed on the whole LA interval. This is actually what occurs in the main detailed search, however, using context-dependent models and applying the true lexical constraints which are not taken into account in this fast match. In practice, this solution is efficiently achieved by exploiting the ergodicity of one-state models. In this case, it simply amounts to perform some "book-keeping" operations on the partial LA scores that are stored in a ring buffer.

5. EXPERIMENTAL RESULTS

This new pruning procedure has been tested in the framework of one-pass 64k trigram decoding using speaker-independent models without any kind of acoustic or linguistic adaptation. All experiments have been made on DEC stations running at 500 MHz using a general purpose research software. The mixture log-likelihood calculations are sped up using density selection methods similar to what is described in [7]. In addition, the 64-bits arithmetic is exploited to compute either four or eight vector components in parallel, this last case implying a slight loss in the accuracy.

5.1. North American Business

This task concerns read speech dictation in quiet office environment. One-pass decoding is performed on the joined dev and eval 1994 sets with a 64k trigram achieving an OOV rate of 0.64%. For the baseline system, the pruning thresholds have been tightly adjusted to get the "nominal" accuracy at the cheapest cost.

The influence of the LA interval length Δ_t has been first investigated. Table 1 presents some typical results obtained on the female portion of the test data (20 speakers, 7770 spoken words). The first four lines relate to the same pruning setup, the only varying parameter being Δ_t while the last line is obtained with a smaller value of the f_{LA} threshold. As expected, the search space reduction measured from the average number of expanded arcs gets larger when increasing the LA interval length up to 9 centiseconds. Looking at the error rates, it appears that the best strategy is to prune somewhat more with a shorter length of 6 instead of 9, leading to a relative degradation by 3.7%. The overall decoding cost is more than halved as indicated by the real-time factor (RTx) figures.

Δ_t Csec	Av. # Arcs (ratio)	RTx	WER %
0	3409 (./1.0)	2.35	9.47%
3	2332 (./1.5)	1.63	9.55%
6	1695 (./2.0)	1.40	9.63%
9	1413 (./2.4)	1.21	9.90%
6'	1362 (./2.5)	1.15	9.82%

Table 1: Influence of LA Interval Length Δ_t

A second point concerns the efficiency of the pruning rule exploiting the anticipated phoneme matches. Phoneme look-ahead is performed on 6 centi-second frames and the acoustic scores are used for pruning either in "cascade" after the linguistic LA scores, or in combination with, as described by the last equations of section 3. As shown in table 2, combining the acoustic and linguistic expectations in a single beam pruning equation is the most robust strategy, achieving the largest search reduction at the least error rate increase. This is due to the complementary nature of the LM and acoustic contributions working in synergy.

LA Pruning	#Arcs	Rtx	WER
Bigram LM only	3409	2.35	9.47%
Cascade LM, AC	2087	1.48	9.99%
Combined LM & AC	1695	1.40	9.63%

Table 2: Comparison of several LA pruning rules

To get more insight into the overall speed-up problem, table 3 presents the breakdown of the decoding costs in terms of search effort, log-likelihood computations and fast acoustic match. The computations involved by the bigram smearing process are included in the "Search" part while the remaining costs for the trigram LM are quite small due to a cache, between 2 and 3% of the whole CPU.
The figures in table 3 have been computed on the male portion of the test sets (20 speakers, 7803 spoken words) for 3 different pruning setups : a "safe" mode without significant loss in accuracy, a fast mode with a relative loss of 5% and an "aggressive" pruning setup implying a degradation of 19%, from 10.5 to 12.5% error rate. The computing resources are given in terms of CPU units.

mode	CPU	#Arcs	Search	Ac.Lik.	FM
base	6180	3944	56.5%	41.5%	0.0%
safe	3955	2001	47.5%	48.5%	1.1%
fast	3020	1146	42.5%	54.0%	1.6%
aggr	1897	612	27.1%	67.5%	2.4%

Table 3: Breakdown of decoding cost in main parts

In the baseline decoder, the "pure" search costs are dominant while in the other cases the acoustic likelihoods become gradually the most intensive CPU part. In all cases, the acoustic fast match represents an almost negligible burden. Compared to the baseline, the search effort of the "aggressive" mode is reduced by a factor of 7, from 3490 to 514 units.

Table 4 presents the evaluation results over the whole test set for three different pruning setups and shows the gain in

decoding efficiency achieved by the combined LA pruning method. Compared to the baseline, the error rate for real-time decoding is improved by 1% absolute due to a better focus of the explored search space.

THR	Linguistic LA only			Acoust. & Linguist. LA		
	#Arcs	RTx	WER	#Arcs	RTx	WER
265	3500	3.1	10.0%	2100	1.9	10.0%
250	1600	1.7	10.9%	1100	1.3	10.6%
240	1100	1.1	12.1%	750	0.9	11.1%

Table 4: Evaluation on NAB : 40 speakers, 15573 words

5.2. US Broadcast News (Hub4)

The same algorithms have been evaluated on broadcast news recordings, again in the context of a (first) one-pass 64k trigram decoding. The test data consists of the hand partitioned Hub4 evaluation set of 1997 comprising about 3 hours with 32832 spoken words. Table 5 contains figures about the size of the search space that clearly show the difficulty of this multi-style and multi-channel task as opposed to read dictation using a single microphone.

LA-Pruning	#Arcs	Del-Ins-Sub	WER %	Rtx
2G-LM only	28395	4.52 2.16 14.37	21.05%	17.3
AC & 2G-LM	12353	4.56 2.17 14.40	21.12%	9.5
AC'& 2G-LM	5528	4.73 2.21 15.54	22.49%	3.3

Table 5: Combined Ac. & Ling. LA on Hub4 Eval'97

As shown in table 5, the combined acoustic and linguistic LA pruning reduces the average number of expanded arcs by a factor of 2.3 with little degradation in the accuracy (0.4% relative). The overall decoding cost is reduced by a factor of 1.8 due to the fact that the log-likelihood computations are not reduced in the same proportion as the search space. The last line of the table presents results obtained with a tighter setup for the likelihood computations as well as with a more aggressive pruning, showing a reduction of the search space and the real-time factor by about 5, however, at the price of an error rate increased by 6.9% relative, up to 22.5%.

These experiments are summarized in the two curves shown below, which give the word error rate in terms of the overall real-time factor for the baseline and the improved decoder over a wide range of operating-conditions, using the same set of acoustic models. The advantages of combined look-ahead pruning are clear even though the curve slope remains quite steep below two times real-time.

6. CONCLUSION

Although there is an obvious tradeoff between speed and accuracy, the reduction of the search space appears quite appreciable just by looking ahead of one phonetic arc and less than one-tenth of a second of speech. Further improvements are expected by replacing the context-independent coarse models with left diphones as it has been done in [5], to better model the across-phone coarticulation and to capture some short-range lexical constraints.

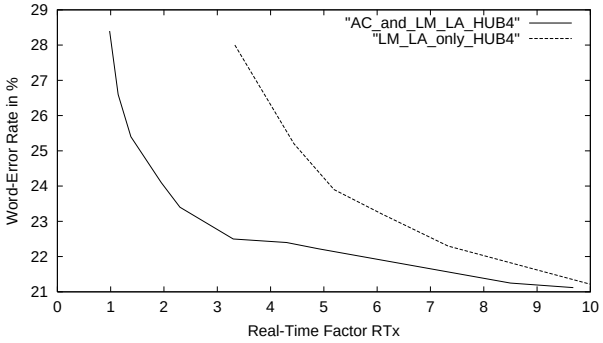


Figure 1: Accuracy vs Real-Time Factor for Hub4.

7. REFERENCES

1. X. L. Aubert, "One Pass Cross Word Decoding For Large Vocabularies Based On A Lexical Tree Search Organization", pp. 1559–1562, Proc. EUROSPEECH, Budapest, Hungary, Sep. 1999.
2. Ney, H., Haeb-Umbach, R., Tran, B.-H. & Oerder M., "Improvements in beam search for 10000-word continuous speech recognition", pp. 13–16 in Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, San Francisco, CA, March 1992.
3. Stefan Ortmanns, Andreas Eiden, Hermann Ney and Norbert Coenen, "Look-Ahead Techniques for Fast Beam Search", pp. 1783–1786, in Proc. ICASSP'97, Munich, Germany, 1997.
4. Fil Alleva, "Search Organization in the Whisper Continuous Speech Recognition System", in Proceedings of the IEEE ASRU Workshop, pp 295–302, Santa Barbara, California, USA, Dec. 1997.
5. M. Novak and M. Picheny, "Speed Improvement of the Time-Asynchronous Acoustic Fast Match", pp. 1115–1118, Proc. Eurospeech'99, Budapest, Hungary, Sep. 1999.
6. Long Nguyen and Richard Schwartz, "The BBN Single-Phonetic-Tree Fast-Match Algorithm", pp. 1827–1830, Proc. ICSLP, Sydney, Australia, Nov. 98.
7. Stefan Ortmanns, Thorsten Firzlaff and Hermann Ney, "Fast Likelihood Computation Methods for Continuous Mixture Densities in Large Vocabulary Speech Recognition", pp. 139–142 in Proc. Eurospeech'97, Rhodes, Greece, Sep. 1997.