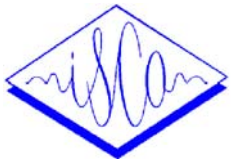

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A FAST VERSION OF THE ATROS SYSTEM*

M. J. Castro(1) *D. Llorens*(2) *J.A. Sánchez*(1) *F. Casacuberta*(1) *P. Aibar*(2) *E. Segarra*(1)

(1) Depto. de Sist. Informáticos y Computación
Universitat Politècnica de València
Cami de Vera s/n
46022 València (Spain)

(2) Unidad Predepartamental de Informática
Universitat Jaume I
Campus Riu Sec
12071 Castelló (Spain)

Abstract

ATROS is an automatic speech recognition/understanding/translation system whose knowledge sources (acoustic models, lexical models, syntactic language models, semantic models and translation models) can be learnt automatically from training data by using similar techniques. The search process in ATROS is performed through a *Synchronous Beam Search* technique.

In this paper, a faster version of ATROS is presented and evaluated. This version supports improved acoustic and syntactical models. It also incorporates improved search algorithms to reduce and the computational requirements for decoding: *Fast Phoneme Look-Ahead* and *Histogram Pruning*. The system has been tested on a Spanish task of queries to a geographical database (with a vocabulary of 1,264 words). The best result achieved (in real time) was 7.10% of word error rate.

10.21437/Eurospeech.1999-317

1 System overview

Optimal speech decoding based on a search process in an integrated network of different knowledge sources is a hard computational problem [1]. The Viterbi decoding is an efficient approach, but for real applications the required networks are very large, and the corresponding search process requires a lot of computational effort to achieve good performances. Developing efficient internal representations of the knowledge sources and efficient search algorithms are crucial to fast decode speech utterances.

ATROS (Automatically Trainable Recognizer Of Speech) is an automatic speech recognizer that can be used for speech decoding into text, for speech understanding and for speech translation. The main characteristic of the system is the possibility that all knowledge sources (acoustic models, lexical models, syntactic language models, semantic models and translation models) can be learnt automatically from real data [2, 3, 4, 5, 6].

ATROS is composed of two parts: the *feature extraction module* and the *decoding module*. The first module computes a sequence of feature vectors from the input speech signal. From this sequence, the second module computes an output in a search process on an implicit finite-state network that represents the integration of all the knowledge sources.

Typically, the feature extraction module produces a feature vector of 33 components (10 mel-cepstrum coefficients plus the energy and their first and second derivatives) every 10 msec.

As acoustic sublexical models, ATROS supports Continuous Density Hidden Markov Models (CDHMMs), which have been trained with the HTK toolkit [7].

The lexical models are represented by stochastic finite-state networks whose transitions are labelled with phone-like models. The corresponding acoustic lexical models consist of word acoustic models which are obtained by the concatenation of acoustic sublexical models according to orthographic-phonetic rules.

General stochastic finite-state automata and n-grams can be used as syntactic language models. In both cases, the models can be learnt automatically from training sentences [8, 9], and can be smoothed with other n-grams. For speech understanding and speech translation, stochastic finite-state transducers are used. These models can be also learnt automatically from training pairs [3, 4, 5]. In any case, the acoustic lexical models are integrated in the finite-state structures that represent the input syntactic constraints modeled in the syntactic language models or in the finite-state semantic model or in the finite-state translation model.

For speech decoding, syntactic language models (n-grams or stochastic finite-state automata) are used and the decoding process returns the optimal sequence of words (sequence of lexical transition labels of the syntactic language model). For speech understanding and speech translation, finite-state transducers models are used and the search process returns the optimal sequence of meaning labels, in the first case, or the optimal sequence of output words (the output transition labels of the finite-state transducer associated to the most likely word sequence) in the second case.

The search for the most likely word sequence is approximated by the most likely state sequence in a net-

*Work partially supported by the Spanish CICYT under contract TIC98/0423-CO6.

work that integrates the acoustic, lexical and syntactic (semantic or translation) models. This search process is performed by using the Viterbi algorithm together with an heuristic for pruning the less likely histories. This search strategy is known as *Synchronous Beam Search* [10].

All word models that correspond to transitions which leave from a particular state of the language model are represented as a prefix-tree (tree lexicon) [11]. The use of this internal representation in the Synchronous Beam Search and the possibility of using the language model probabilities in the tree as soon as possible (*Language Model Look-Ahead*) let to reduce the computational search time required without decreasing the system performance [2, 6].

The rest of the paper is organized as follows. Section 2 is devoted to the definition and training the acoustic models and language models respectively. The improved techniques to speed-up the search process are described in Section 3. Section 4 shows the experiments that have been carried out. Finally, some conclusions are mentioned in Section 5.

2 Acoustical, Lexical and Language models

2.1 Acoustical and Lexical models

Context-independent phones were modeled through CDHMMs. The emission probability of each state is represented by a Gaussian mixture density with diagonal covariance matrix. ATROS compute the emission probability density values at each state as the highest probability density value from all of the Gaussian emission of the mixtures. This type of computation allows us to use minus-log values of the probabilities and probability densities, and consequently, the computation of a maximum operator presents a lower computational cost than the addition operator. Each model had three states without skip transitions.

The lexical models are composed by the concatenation of sublexical models to form word acoustic models.

The acoustic models were trained with the following acoustic material: the overall training database gathers 1,529 utterances from 57 speakers (which accounts for nearly 470,000 acoustic frames and 55,000 phonetic units).

The acoustic CDHMM were trained by using the Baum-Welch algorithm of the HTK toolkit [7] from training data parametrized into sequences of cepstral coefficients and energy (and their first and second derivatives) by the ATROS system. Context-independent phone-like units were used, 28 units in total (including initial, middle and final silences), which were defined previously. Models of monophones with 16, 32 and 64 mixtures per state were evaluated. Table 1 shows the total number of mixtures for each type of model.

Table 1: Number of total mixtures for phone-like units.

Sublexical units	Number of units	Maximum mixtures per state	Total number of mixtures
phones	27	16	1,344
phones	27	32	2,687
phones	27	64	5,362

2.2 Language model

A trigram model was estimated with the second version of the Stochastic Language Model Toolkit [12]. The training set used for the estimation of the language model consisted of 8,221 written sentences (78,200 Words) of Queries to a Spanish Geographic information Database (GDQ) [13], with a vocabulary of 1,264 words. A test set of 1,138 different written sentences (11,200 Words) was used to measure the perplexity of the obtained model. The perplexity of the test set with the trigram model was 10.22.

3 Speed-up techniques

3.1 Fast Phoneme Look-Ahead

The *Fast Phoneme Look-Ahead technique* [14] has been incorporated in ATROS in order to reduce the number of hypotheses which are considered in the search process and, consequently, to reduce the search time. The main idea of the Fast Phoneme Look-Ahead consists of determining whether every new phoneme model which is going to be started is likely to survive pruning steps in the future. This is decided by computing an approximate score (*look-ahead score*) using a simpler phoneme model (*look-ahead model*) and some future time frames (*look-ahead buffer*). Several hypotheses could continue with the same phoneme model and therefore this computation should only have to be carried out once. The look-ahead score is combined with the exact score of the predecessor phoneme model and the phoneme is started only if this new value is over a certain threshold (*look-ahead threshold*) in way similar to the beam search. If the phoneme model is started, then the exact score is computed. This means that the optimal path can be pruned and only a suboptimal solution may be achieved.

The fast look-ahead scores are computed for every time frame not by using the exact phoneme models, but rather by using a simpler one in order to reduce the amount of computation. In ATROS, the

look-ahead models had three states (the same as the exact phoneme models) and a few densities in each state. These models were also trained by using the HTK toolkit.

The performance of the system depends on how well all the tuning parameters are adjusted. In the current system this adjustment is made by comprehensive experimental work [15].

3.2 Histogram Pruning

In the Viterbi beam search approach, only the hypotheses whose scores are relatively close to the best hypothesis are considered. The beam width is fixed through a predefined pruning threshold.

In the decoding experiments, peaks of active hypotheses that were several orders of magnitude higher than the average number of active hypotheses were observed. *Histogram Pruning* [16] is a technique that allows setting an upper limit to the number of active hypotheses. By using a histogram of the hypotheses scores, the pruning threshold could be decreased in order to keep the number of active hypotheses below this limit.

We introduce this technique through an efficient implementation with negligible overhead. This implementation allowed us to obviate a second pass through the data structure that keeps the active hypotheses. The additional pruning of the active hypotheses in a frame is carried out in the next input frame: only hypotheses whose scores were over the previous histogram pruning threshold were analyzed.

4 Evaluation of the fast version of the system

In this section, we present some experiments that were carried out to evaluate the performance of the two fast versions of the system. The task was the GDQ (previously described) with a vocabulary of 1,264 words.

The performance of the system was measured on a test set which consisted of 600 utterances from 12 speakers (200 different sentences, 5,655 words) from the GDQ application task [13]. Note that the GDQ database and the utterances used to train the acoustic models were independent with respect to the speakers, the text and the task.

To evaluate the performance of the system, we matched each decoded utterance against the correct transcription of the sentence (in terms of a sequence of words). Then, the word-error rate (*wer*) was calculated.

For each experiment, we show the obtained word-error rate and a measure of the consumed time, given by the number of seconds which were necessary to process one hundred frames (equivalent to approximately real time). All the experiments were performed on a SGI2 workstation R10000 with 384 MB of RAM.

Different beam-search and grammar-scale factors were proved and the best results for each type of acoustic unit without any speed-up technique and with Fast-Phoneme Look-Ahead and Histogram Pruning techniques are shown in Table 2. As can be seen from the obtained results, better performance was obtained using the 32-gaussian models. It can be pointed out that Histogram Pruning clearly outperforms the Fast-Phoneme Look-Ahead technique in all cases (with a *wer* of 7.10% versus 8.96% in real time with the 32-gaussian models).

Table 2: The word-error rate (*wer*) obtained for the test set along with the real time factor (*rtf*) is shown. The sublexical units and the maximum number of mixtures per state are shown in the first and second column, and the speed-up technique in the third column (none, FLA: fast look-ahead, HP: histogram pruning).

Sublexical units		Speed-up technique	*wer*	*rtf*
phones	(16)	–	8.00	48.5
phones	(16)	FLA	10.27	0.8
phones	(16)	HP	8.23	0.7
phones	(32)	–	6.93	49.4
phones	(32)	FLA	8.96	1.1
phones	(32)	HP	7.10	1.2
phones	(64)	–	7.08	48.3
phones	(64)	FLA	9.65	1.8
phones	(64)	HP	7.33	2.0

5 Summary

In conclusion, both Fast-Phoneme Look-Ahead and Histogram Pruning in the ATROS system have allowed us to significantly reduce the search effort without producing a significant increase in word error rate. Consequently, these techniques allow us to use more powerful and computationally expensive models and/or wider beams. Histogram Pruning, in particular, works much better than Fast-Phoneme Look-Ahead. The best result achieved (in real time) with this speed-up technique was 7.10% of the word error rate for the GDQ task.

Finally, we expect to significantly improve the performance of the system by using both speed-up techniques together. In addition, we are training triphones by using decision-tree techniques in order to improve the acoustic models. We also plan to use a

category-based trigram language model (that is, clustering the words into categories by domain knowledge; for the GDQ task, categories such as rivers, mountains, seas, etc) [13].

References

[1] F. Casacuberta and C. de la Higuera. Linguistic Decoding is a Difficult Computational problem. *Pattern Recognition Letters*, 1999. To appear.

[2] D. Llorens, V.M. Jimenez, J.A. Snchez, E. Vidal, and H. Rulot. ATROS, an Automatically Trainable Continuous-Speech Recognition System for Limited-Domain Tasks. In *VI Spanish Symposium on Pattern Recognition and Image Analysis*, pages 478–483, Córdoba (Spain), 1995.

[3] V. M. Jiménez, A. Castellanos, and E. Vidal. Some Results with a Trainable Speech Translation and Understanding System. In *Proceedings of the ICASSP'95*, pages 113–116, Detroit, MI (USA), May 1995.

[4] J. C. Amengual et al. Speech Translation Based on Automatically Trainable Finite-State Models. In *Proceedings of the Eurospeech'97*, volume 3, pages 1439–1442, Rhodes (Greece), 1997.

[5] E. Vidal. Finite-State Speech-to-Speech Translation. In *Proceedings of the ICASSP'97*, volume 1, pages 111–114, Munich (Germany), 1997.

[6] D. Llorens, F. Casacuberta, E. Segarra, J.A. Sánchez, P. Aibar, and M.J. Castro. Acoustic and syntactical modeling with the ATROS system. In *Proceedings of the ICASSP'99*, pages 641–644, Phoenix (Arizona), USA, March 1999.

[7] S.J. Young, P. C. Woodland, and W.J. Byrne. HTK: Hidden Markov Model Toolkit V1.5. Technical report, Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc., 1993.

[8] E. Segarra and L. Hurtado. Construction of Language Models using the Morphic Generator Grammatical Inference (MGGI) Methodology. In *Proceedings of the Eurospeech'97*, pages 2695–2698, Rhodes (Greece), 1997.

[9] P. Clarkson and R. Rosenfeld. Statistical Language Modeling using the CMU-Cambridge toolkit. In *Proceedings of the Eurospeech'97*, pages 2707–2711, Rhodes (Greece), 1997.

[10] H. Ney, D. Mergel, A. Noll, and A. Paeseler. Data Driven Search Organization for continuous Speech Recognition. *IEEE Transactions on Signal Processing*, 40(2):272–281, 1992.

[11] H. Ney, R. Haeb-Umbach, B.-H. Tran, and M. Oerder. Improvements in beam search for 10000-word continuous A speech recognition. In *Proceedings of the ICASSP'92*, volume 1, pages 9–12, San Francisco, California (USA), March 1992.

[12] R. Rosenfeld. The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation. In *ARPA Spoken Language Technology Workshop*, Austin, Texas (USA), 1995.

[13] J. E. Díaz-Verdejo, A. M. Peinado, A. J. Rubio, E. Segarra, N. Prieto, and F. Casacuberta. ALBAYZIN: a Task-Oriented Spanish Speech Corpus. In *First International Conference on Language Resources and Evaluation*, pages 497–591, Granada (Spain), 1998.

[14] S. Ortmanns, A. Eiden, H. Ney, and N. Coenen. Look-ahead techniques for fast beam search. In *Proceedings of the ICASSP'97*, pages 1783–1786, Munich (Germany), 1997.

[15] J.A. Snchez, F. Casacuberta, P. Aibar, D. Llorens, and M.J. Castro. Fast phoneme look-ahead in the atros system. 1999. Accepted in VIII Spanish Symposium on Pattern Recognition and Image Analysis.

[16] V. Steinbiss, B.-H. Tran, and H. Ney. Improvements in Beam Search. In *Proceedings of the ICSLP'94*, pages 2143–2146, Yokohama (Japan), September 1994.