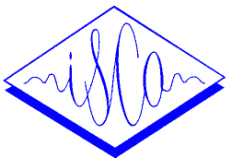

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

# SENONES, MULTI-PASS SEARCH, AND UNIFIED STOCHASTIC MODELING IN SPHINX-II

M.Y. Hwang F. Alleva X. Huang

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 U.S.A.

## ABSTRACT

SPHINX-II is designed for large vocabulary, speaker-independent continuous speech recognition and is based on semi-continuous hidden Markov models. In the November 1992 ARPA speech evaluation, SPHINX-II achieved the lowest error rate (5%). This paper concentrates on the special techniques that made SPHINX-II successful and different from other systems. Specifically these include senonic decision trees for acoustic modeling, the multi-pass decoder to meet the challenge for very large vocabulary recognition, and the unified stochastic engine for jointly optimizing the acoustic and language model.

Keywords: Shared-distribution models, senones, decision trees, multi-pass decoder, unified stochastic engine.



## 1 INTRODUCTION

This paper describes our solutions to the following three problems: detailed acoustic modeling for unseen triphones, an efficient and accurate search algorithm for very large vocabulary tasks, and joint optimization of acoustic and language models.

Unseen triphones often occur on large vocabulary tasks for we have limited training data. They are often replaced by more general models like the monophone or biphone models. To alleviate the over-generalization problem, decision trees have been used in many areas [5, 13, 2]. [7] used decision-tree based generalized triphones to provide better models for unseen triphones. However, [10] showed that the shared-distribution model (SDM) offered more accurate modeling than the agglomerative generalized triphone [11]. The goal of the SDM was to search for similar Markov states so that their output distributions can be averaged to form a *senone*, which is a general and still *accurate* representation of the original component distributions. To combine both advantages of the decision tree and the SDM, this paper presents the decision-tree based senone.

We automatically built a decision tree, $tree(P, s)$, for each state $s$ of each phone $P$ to classify all the $s$-th Markov states of all $P$-triphones. Each leaf is a cluster of similar states, which define a senone. Given a $P$-triphone, seen or unseen in the training data, we traverse $tree(P, s)$ until a leaf is reached to find the senone to be associated with state $s$ of that triphone. Predicting unseen triphones using senones reduced the word error rate by about 10% compared with using context-independent models for unseen triphones. Note the same set of senones represents both seen and unseen triphones; no additional parameters are needed to model unseen triphones.

Recent work on search algorithms for continuous speech recognition has focused on three dimensions: large vocabularies, long distance language models and detailed acoustic modeling. To meet this challenge we incrementally apply all available acoustic and linguistic information in three search passes [1]. Pass one is a standard, time-synchronous, left-to-right Viterbi beam search which produces word end times and scores with a bigram language model. Pass two, guided by the results from pass one, is a right-to-left beam search that produces word begin times and scores. Pass three is an $A^*$ search that combines the results of pass one and two with a long distance language model. With this three-pass decomposed incremental search, our objective to maximize recognition accuracy with a minimal increase in computational complexity is satisfactorily fulfilled.

In most speech recognition systems, acoustic and language models are usually constructed separately, where language models are derived from a large text corpus without considering confusable acoustic data, and acoustic models are optimized without considering language model discrimination capacity. The unified stochastic engine (USE) is a general framework to jointly optimize both the acoustic model and language model [8]. However, for our preliminary study we applied USE only to language-weight learning. With about 1000 training utterances, these automatically learned word-dependent language weights reduced word error rates by 5% across different test sets.

This paper is organized as follows. The decision-tree based senone is elaborated in Section 2. Section 3 describes the three-pass decoder. USE is explained in Section 4. Section 5 presents the experimental results and analysis.

## 2 The SENONIC DECISION TREE

Unseen triphones occur often on large vocabulary tasks due to the limited amount of training data. The bottom-up agglomerative clustering algorithm for the SDM [10] does not have the capability to decide which distribution clusters the Markov states of an unseen triphone belong to. Therefore, in the past we used context-independent phone models for unseen triphones during decoding. To model unseen triphones with senones, we extend the principle of state sharing to the decision-tree classification [7].

When extending the decision-tree algorithm from triphone classification to Markov state classification, we modified two things,

besides the object that is being classified. First we notice that the computation for the tree classification with composite questions is essentially an exponential function of the number of the input objects, which for state classification is the number of triphones times the number of Markov states per triphone HMM. For a large-vocabulary task, such as the 5,000-word Wall Street Journal (WSJ) task, the input size for the classification procedure grows dramatically. To speed classification, we enforce the state-dependency constraint, which disallows output distributions from different topological locations being in the same cluster. This is informed by the SDM study on the 1,000-word Resource Management task, where we found that states in the same cluster were mostly from the same topological location [10]. In SPHINX-II, there are 50 phonemes and there are 5 states for each phonetic HMM. Therefore, we built automatically 250 decision trees to classify all Markov states in all triphone models, with an algorithm similar to the one used in [7]. At the beginning of the classification, a set of linguistic questions were created manually. Then the tree growing algorithm automatically determines which question is the best for each node. The question associated with a tree node can be simply one of the hand-crafted questions, or a composite question formed automatically by conjunction, disjunction and/or negation of the simple questions.

Secondly, we found that it is helpful to take neighboring state information into consideration for the tree growing. To elaborate the concept, suppose we are classifying all the $s$-th states of all $P$-triphones and we are deciding which question is the best to split node $n$ into two children, $n_1$ and $n_2$. Although state $s$ is the only state under consideration, we keep the output distributions associated with the other states at the same time. To compute the goodness of question $Q$ with respect to state $s$, we first compute the entropy decrease at each state $i$ due to the split:

$$\Delta H_i = C_i(n)H_i(n) - C_i(n_1)H_i(n_1) - C_i(n_2)H_i(n_2)$$

where $H_i(n)$ is the entropy of state $i$ of node $n$, and $C_i(n)$ is the occurrence count of state $i$ of node $n$, which is estimated by the Baum-Welch algorithm [3]. Next the goodness of question Q with respect to state $s$ is a weighted summation of all the entropy decreases, with bigger weights for states closer to $s$:

$$\text{goodness}(Q, s) = \sum_i w_i(s)\, \Delta H_i \qquad (1)$$

The more entropy decreases, the better the question is. Note that exactly the same amount of computation is required for classifying different states of the same phone. The only difference between them is the combining weights $w_i(s)$. When uniform combining weights are adopted,

$$\text{goodness}(Q, s) = \sum_i \Delta H_i \quad \forall s$$

and thus the classification degrades to yield the generalized triphone. On the other hand, when $w_i(s) = 0\ \forall i \neq s$,

$$\text{goodness}(Q, s) = \Delta H_s \quad \forall s$$

information from neighboring states is not used. In other words, (1) becomes a general distance measure which can be derived to either generalized triphones or variants of senonic models.

We stop tree growing when the best question fails to offer a minimum entropy decrease. Each leaf of each tree represents a senone. After the senone mapping trees are created for all seen triphones, we retrain the triphone HMMs using the mapping. During testing, any triphone, either seen or unseen in the training data, traverses the five decision trees corresponding to the phone until leaves are reached, where we find the five senones to be associated with the five states of the given triphone. Thus, we are able to represent both seen and unseen triphones in a unified way by the same set of senones.

Figure 1 illustrates an example tree for classifying the first state of $K$-triphones. It also highlights the path when the /K/ triphone in the word "*welcome*" traverses the tree. The left phoneme /L/ is an element in set *SON* and *BACK-L*. As expected, for the first state of a phone, questions about the left context are more important than those about the right context.

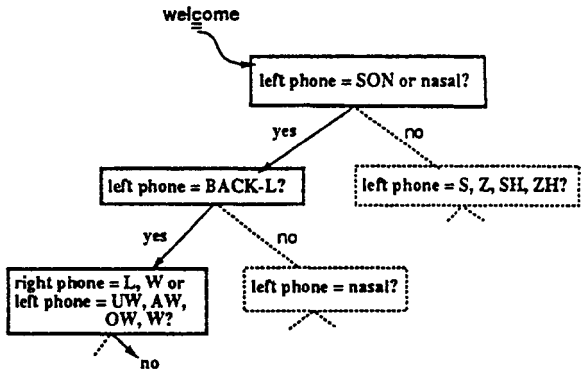


Figure 1: Part of the decision tree for classifying the first state of $K$-triphones.

## 3 THE MULTI-PASS DECODER

Recent work on search algorithms for continuous speech recognition has focused on three dimensions: large vocabularies, long distance language models and detailed acoustic modeling. To meet the challenge, several systems have been proposed to use Viterbi beam search as a fast-match [15, 16] for stack decoding or the $N$-best paradigm [14]. In these systems, simple acoustic and language models are used to produce $N$-best hypotheses. Multi-pass rescoring is subsequently applied to these hypotheses to produce the final recognition output. One problem in this paradigm is that decisions made by the initial pass are based on simplified models. This results in errors that the rescoring procedure cannot recover once the correct hypothesis is not in the $N$-best list, which is very likely when the utterance is very long or the vocabulary is large and contains a lot of confusable words. Specifically, we compared

the word error bound of two decoders, one using between-word triphones, the other using within-word triphones only. The error bound is simply the least errorful of the $N$-best hypotheses produced by the system. Figure 2 plots the error bound for the two systems. Both systems are configured with 7000 senones on the WSJ task. In Figure 2 we see that the error bound for the between-word system is consistently 22% to 25% better than the within-word system.

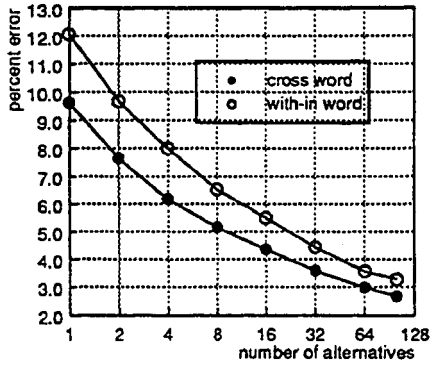


Figure 2: Comparison of a within-word triphone system versus a between-word triphone system. The number of hypotheses considered versus the word error bound is plotted.

Therefore, instead of the $N$-best rescoring paradigm, we designed a three-pass decoder which applies directly the most detailed models in the search. Pass one is a standard left-to-right Viterbi beam search, using the bigram language model and the most detailed acoustic model including between-word triphones. It produces the most likely word sequence for the input utterance. For most applications, the decoder ends here.

However, when long-distance language models are available or when the $N$-best hypotheses are desired (e.g. for a natural-language understanding system), the first pass also produces word end times and scores, and two other passes are invoked. The second pass, guided by the results from the first pass, is a right-to-left Viterbi beam search which produces word beginning times and scores. Pass three is an $A^*$ search which combines the word lattices from pass one and two with a long distance language model to produce the top $N$ hypotheses. Delaying the long-distance language model until the $A^*$ search is much simpler than applying the complex language model directly in the beam search. Moreover, since long-distance language models are used during the $A^*$ search, we can produce significantly better top $N$ hypotheses than using simple language models like the other systems do.

# 4 THE UNIFIED STOCHASTIC ENGINE

The basic idea behind USE is to have an objective function defined by both the acoustic and language models so that minimizing the objective function leads to optimizing jointly both the acoustic and language models. The objective function we defined is related to the goal of maximizing the difference between the correct word sequence $\theta$ and the incorrect one $\bar{\theta}$ ( which can be generated by the $A^*$ search described above) for each individual utterance $\mathbf{x}$ in the training data:

$$\min \ \Delta(\mathbf{x}) = \sum_{\bar{\theta}} p(\bar{\theta}|\mathbf{x})\{score(\bar{\theta}|\mathbf{x}) - score(\theta|\mathbf{x})\}$$

where

$$score(\theta|\mathbf{x}) = \log Pr(\mathbf{x}|\mathrm{HMM}(\theta)) \ + \ \log Pr(\theta, \alpha(\mathbf{x}, \theta)) \quad (2)$$

similarly for $score(\bar{\theta}|\mathbf{x})$. The first term in (2) is the acoustic score and the second term is the language score weighted by an ad-hoc value $\alpha$, which can depend on both the word sequence $\theta$ and the given acoustic data $\mathbf{x}$. In most systems, $\alpha$ is a constant, simplifying the second term to be $\alpha \ \log Pr(\theta)$. The overall objective function for USE is a summation of a normalized $\Delta(\mathbf{x})$ for all $\mathbf{x}$ in the training data:

$$\min \ obj = \sum_{\mathbf{x}} \frac{1}{1 + e^{-\Delta(\mathbf{x})}} \quad (3)$$

By taking the partial derivative of (3) and using the gradient descent algorithm [6], we are able to optimize any interesting parameter or all parameters simultaneously. In our preliminary study we only optimized the language weight, $\alpha$.

# 5 EXPERIMENTAL PERFORMANCE

## 5.1 Baseline System Description

SPHINX-II was evaluated on the ARPA 5,000-word WSJ speaker-independent continuous speech recognition task (non-verbalized punctuation). The standard bigram, with a test-set perplexity of 118, was supplied by MIT Lincoln Lab [12]. There are 7200 sentences from 84 speakers in the official training set and 330 utterances from 8 new speakers in the test set.

SPHINX-II front-end consists of 4 sets of independent features, including power and mel-frequency cepstral coefficients normalized by the sentence-based cepstral mean, and 1st-order and 2nd-order differences. Sex-dependent semi-continuous HMMs [9, 4], each gender with 7000 decision-tree based senones, are trained independently by the Baum-Welch algorithm. The 5-state Bakis topology is used for all phonetic HMMs.

## 5.2 Evaluation Results

For each test utterance, we ran only the first pass of the decoder on both the male and female models in parallel and output the word sequence that had a better score. The word error rate was 7.4%, as shown in the first row of Table 1.

To avoid dramatic degradation on outlier speakers, we added the generic model (male mixed with female) into the parallel decoder. The resulting error rate of 7.3% showed that there was no outlier speaker in this test set.

| system | error rate | reduction |
|---|---|---|
| Female/Male | 7.4% | – |
| Female/Male/Generic | 7.3% | 1% |
| +USE | 6.9% | 5% |
| +Trigram | 5.3% | 23% |

Table 1: Word error rates on the 5K-nvp WSJ task.

The above experiments were run under a constant language weight which was tuned on a separate development set. To conduct experiments for USE, we used the three-pass decoder to generate the top 100 hypotheses for each utterance in the training set. We then included only those utterances whose correct word sequences were among their top 100 hypotheses. Due to the small amount of training data (about 1000 utterances), we assumed $\alpha(\mathbf{x}, \theta)$ was independent of the acoustic data $\mathbf{x}$ and dependent only on the last word for all partial theories of $\theta$. We used the gradient descent method based on the 1000 utterances to learn these word-dependent language weights. The learned word-dependent language weights have not yet been incorporated into the decoder. Therefore we used the paradigm of rescoring the $N$-best hypotheses instead and it gave us only 5% error reduction, mainly because of insufficient training data.

Due to time constraints, we were not able to incorporate the trigram language model into the $A^*$ search of the decoder. Instead, we again rescored the top 100 hypotheses which were generated with the bigram language model. Trigram rescoring dropped the word error rate to 5.3%. The fact that the 23% error reduction was smaller than what we expect from trigram was explained by the fact that for some utterances, the correct word sequences were not covered in the top 100 hypotheses. As the utterance gets longer and/or the vocabulary gets larger, we will have to generate exponential $N$ hypotheses in order to cover the correct word sequence. This again fails the $N$-best paradigm and supports our belief of applying all knowledge sources at the early stage.

## 6 Conclusion and Future Work

Senones offer accurate and detailed acoustic modeling. The multi-pass decoder maximizes recognition accuracy by applying the most detailed acoustic model in the early phases of the decoder with a minimal increase in computational complexity. The unified stochastic engine gives us a general framework for optimizing all parameters in a speech recognition system.

Our short-term goal is to embed the USE learned language weights into the first two passes of the decoder and incorporate the long-distance language model into the $A^*$ search. The long-term goal is to apply USE to other parameter optimizations, e.g. acoustic corrective training.

## Acknowledgements

This research is sponsored by ARPA Order 7239, under contract N00039-91-C-0158. The authors would like to express their gratitude to Professor Raj Reddy's encouragement and support, and other members of the CMU speech group for their help.

## References

[1] Alleva, F., Huang, X., and Hwang, M. *An Improved Search Algorithm for Continuous Speech Recognition.* in: IEEE International Conference on Acoustics, Speech, and Signal Processing. 1993.

[2] Bahl, L., Brown, P., de Souze, P., and Mercer, R. *A Tree-Based Statistical Language Model for Natural Language Speech Recognition.* IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-37 (1989), pp. 1001–1008.

[3] Baum, L. E. *An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes.* Inequalities, vol. 3 (1972), pp. 1–8.

[4] Bellegarda, J. and Nahamoo, D. *Tied Mixture Continuous Parameter Models for Large Vocabulary Isolated Speech Recognition.* in: IEEE International Conference on Acoustics, Speech, and Signal Processing. 1989, pp. 13–16.

[5] Breiman, L., Friedman, J., Olshen, R., and Stone, C. Classification and Regression Trees. Wadsworth, Inc., Belmont, CA., 1984.

[6] Gill, P., Murray, W., and Wright, M. Practical Optimization. Academic Press, 1981.

[7] Hon, H. and Lee, K. *CMU Robust Vocabulary-Independent Speech Recognition System.* in: IEEE International Conference on Acoustics, Speech, and Signal Processing. Toronto, Ontario, CANADA, 1991, pp. 889–892.

[8] Huang, X., Belin, M., Alleva, F., and Hwang, M. *Unified Stochastic Engine (USE) for Speech Recognition.* in: IEEE International Conference on Acoustics, Speech, and Signal Processing. 1993.

[9] Huang, X. and Jack, M. *Semi-Continuous Hidden Markov Models with Maximum Likelihood Vector Quantization.* in: IEEE Workshop on Speech Recognition. 1988.

[10] Hwang, M. and Huang, X. *Shared-Distribution Hidden Markov Models for Speech Recognition.* IEEE Transactions on Speech and Audio Processing, vol. 1 (1993).

[11] Lee, K. *Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition.* IEEE Transactions on Acoustics, Speech, and Signal Processing, April 1990, pp. 599–609.

[12] Paul, D. and Baker, J. *The Design for the Wall Street Journal-based CSR Corpus.* in: DARPA Speech and Language Workshop. Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[13] Sagayama, S. *Phoneme Environment Clustering for Speech Recognition.* in: IEEE International Conference on Acoustics, Speech, and Signal Processing. 1989.

[14] Schwartz, R., Austin, S., Kubala, F., and Makhoul, J. *New Uses for the N-Best Sentence Hypotheses Within the Byblos Speech Recognition System.* in: IEEE International Conference on Acoustics, Speech, and Signal Processing. 1992, pp. 1–4.

[15] Soong, F. and Huang, E. *A Tree-Trellis Based Fast Search for Finding the N-Best Sentence Hypothesis.* in: DARPA Speech and Language Workshop. 1990.

[16] Zue, V., Glass, M., Goodine, Leung, McCandless, Philips, M., Polifroni, and Seneff, S. *Recent Progress on the Voyager System.* in: DARPA Speech and Language Workshop. 1990.