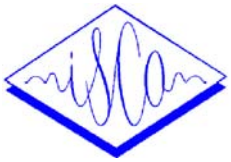

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

CLUSTERING BEYOND PHONEME CONTEXTS FOR SPEECH RECOGNITION

Clark Z. Lee and Douglas O'Shaughnessy
INRS-Télécommunications
16 Place du Commerce, Verdun, Québec, Canada H3E 1H6
E-mail: {clark, dougo}@inrs-telecom.uquebec.ca

ABSTRACT

The clustering of using decision trees is generalized to take into account high-level knowledge sources to better model the co-articulation effects in large vocabulary continuous speech recognition. VQ models are used to reduce the computational cost in constructing decision trees. The search algorithm is designed such that it can provide a general type of information for decision trees without compromising the speed. Experiments with a 30k-word dictionary on the WSJ task show that the word error rate can be reduced by considering additional knowledge sources.

1 INTRODUCTION

In speech recognition, decision trees have been successfully applied to cluster phoneme contexts [1, 2, 3] and HMM states [4, 5, 6], such as quintphone clustering, which has been shown to improve recognition accuracy over triphone models [7]. Since the co-articulation effect in continuous speech is not only caused by neighboring phonemes but is also influenced by other acoustic-phonetic phenomena, notably stresses, syllable and word boundaries, function words, tones, prosodies and so on, presumably we may improve recognition performance by also considering those knowledge sources. Potential problems that could occur if we go beyond phoneme contexts and HMM states for decision trees are that firstly it could be computationally expensive to build the decision trees and secondly it could complicate the search. In this paper, we present our approach to tackle these problems, which allows us to essentially use much more complex acoustic-phonetic models without compromising the speed in our system. Experiments on the *Wall Street Journal* task show that it may increase the recognition accuracy to use decision trees with additional knowledge sources.

The search strategy in our speech recognition system uses two passes [8, 9]. Simple models are used in the first pass to build the word graph, whereas more complex models are used in the second pass to rescore the word graph. Speech signals are processed block by block such that the overall complexity is independent of the length of an utterance or a file.

Since simple models (right-context models) are sufficient to ensure reasonable word search errors in the word graph of the first pass, we use wider phoneme contexts as well as other knowledge sources in the second pass. We introduce various mechanisms to retain different knowledge sources and then pass them to decision trees. Since the computationally expensive part in the second pass is likelihood evaluations, if we maintain about the same number of Gaussian distributions, the speed of the recognizer is essentially independent of the content of the knowledge sources being used.

2 THE DECISION TREES

In order to construct decision trees, we firstly associate each segment in the training data with phoneme contexts, HMM state number, and the answers to whether or not it is a segment of the beginning or ending of a word, whether or not it is a segment of a function word and so on (we distinguish the beginning of a word from the ending of a word rather than only using the union of the two, which is the

cross-word). The decision trees are binary trees in which each node has a list of binary questions and a set of frames associated with it. A node is split into two child nodes according to the maximal gain of the likelihood among the list of questions. Minimal observations for each node and minimal gain of splitting are also imposed. All questions are phonetically and linguistically motivated. Since the decision trees are data-driven, the degree of importance of the questions being asked can be judged by the decision trees.

In principle, a single tree may be used and the question list would include a question about the phoneme itself. Since different phonemes are less likely to be tied together for a reasonable amount of training data, we use one tree per phoneme. However we found the distributions given by different states may have similarities in some cases, so we add the questions about the state numbers.

As the number of binary questions increases, the construction of the decision trees could be quite expensive computationally since we have to calculate the likelihood for each possible splitting on all questions attached to that node. Since the VQ models derived from the tied-mixture models have almost the same performance as the tied-mixture models [10] and also discrete models are much cheaper computationally than continuous models [6], we use the derived VQ models. The VQ models are obtained as follows. We firstly train context-independent continuous models with one distribution per phoneme. Each distribution has its own 256 means but shares the covariance matrix with others. Then we reestimate the distribution weights in a discrete manner by a single iteration of reestimation.

The advantages of using the VQ models are as follows. Firstly since the models are treated as discrete models, it is very efficient for likelihood calculations. Secondly since they are context-independent, we may use them to build the decision trees without iterations. If we label each segment with all necessary information, the decision trees can be constructed with different knowledge sources by simply turning off some questions in the question list.

3 THE SECOND PASS

In order to use the decision trees with more information, we have to retain all information necessary in the second pass. There are three levels of information, phoneme-context level, phonetic transcription level and word level. For the phoneme-context information, we keep track of the look-behind and look-ahead phoneme strings. For phonetic transcription information, such as stresses, syllable and word boundaries, we encode it in the dictionary. For the word information, such as function words and foreign words, we maintain a list and it can be easily accessed by a table look-up.

The objective of the second pass is to find the highest scoring recognition hypothesis by searching the word graph produced by the first pass. We allow for multiple segmentation hypotheses in order to score partial transcriptions exactly, whereas in the first pass, each partial transcription that is hypothesized has a unique segmentation associated with it. We use the depth first search algorithm together with other techniques such as branch ordering, merging and envelope pruning [9] in the second pass.

The principal operation in the second pass is to rescore a branch of the word graph produced by the first pass. A node of the word graph is specified by a quadruple $(\bar{t}, \mathbf{D}, n, \bar{\sigma})$, where $\bar{t}$ is a first pass end time, $\mathbf{D}$ is a look-ahead phone string, n is a node in the lexical graph and $\bar{\sigma}$ is a coarse language model state. A branch b is labeled by a pair $(w, \mathbf{F})$ where w is a word and $\mathbf{F}$ a transcription of w. A second pass partial recognition hypothesis is essentially a partial path in the word graph together with the information needed to support scoring with the fine language models and the fine acoustic phonetic models (the decision trees). The partial recognition hypothesis includes the information $(b, \mathbf{E}, \{\alpha_t\}, \{\Lambda_\sigma\})$ where

(i) b is a branch in the word graph;

(ii) $\mathbf{E}$ is a look-behind phone string;

(iii) $\{\alpha_t\}$ is an array of forward scores centered on $\bar{t}$ whose width is controlled by the uncertainty Δ;

(iv) $\{\Lambda_\sigma\}$ is an array of language model scores indexed by fine language model states σ.

Since b contains word identity w and its phonetic transcription $\mathbf{F}$, we may obtain the phonetic transcription information from $\mathbf{F}$ and word level information by a table look-up from w. The phoneme-context information is from concatenated phoneme string $\mathbf{EF}$. To score a branch, we essentially propagate the array of language scores $\{\Lambda_\sigma\}$ and the array of acoustic scores $\{\alpha_t\}$ for each phoneme-in-context of that branch, where the likelihood of each frame is calculated by passing the phoneme-in-context to the decision trees.

4 EXPERIMENTAL RESULTS

We used the *Wall Street Journal*-based speaker independent CSR corpus with the SI284 training set to train gender-dependent acoustic models. The models were three-state HMMs without skip transitions. Acoustic features were calculated every 10 ms from the 16 kHz sampled data after DC-component removal. The feature vector consisted of 15 cepstral coefficients, 15 delta and 15 delta delta coefficients, where a simple mean normalization was imposed on a fixed window basis. In the first pass and also in building the decision trees, we used the full 45-dimensional feature vector. However we did not use the delta delta coefficients for the second pass models since there was no gain of performance by using them.

The language models were derived from the statistics of *North American Business* texts provided by CMU. The vocabulary was chosen such that the most frequent 30,000 words according to the unigram statistics intersected with the COMLEX dictionary, which resulted in 29,533 words. With this vocabulary, we obtained 5,479,328 bigrams and 6,313,277 trigrams where count 1 statistics were excluded for bigrams and count 3 and below were excluded for trigrams. We used bigrams in the first pass and trigrams in the second pass.

In order to compare the effects of using different knowledge sources for decision trees in the second pass, we fixed the first pass acoustic models, namely decision tree-based VQ models with 3 codebooks and with only right contexts. Each codebook consisted of one covariance matrix, 256 means and a set of distributions. We first trained the tied mixture models with decision trees by clustering the right contexts and then derived the VQ models by reestimating the mixture weights once [10]. For the second pass, the acoustic models had 2 codebooks, each of which had one grand covariance matrix and up to 16 means per distribution. The distributions were selected by decision trees where the minimal observations and minimal gain of splitting were controlled such that the total number of distributions was about the same for using different levels of knowledge sources. We used quintphone contexts as the baseline system. Then we carried out the experiments with additional information (with word boundaries and function words). There were about 330 questions in the decision trees. Question examples are: "is the segment associated with a vowel in its second left phoneme context?" "is the segment associated with the first state in its HMM state context?" "is the segment associated with a function word in its word context?" Since we had gender-dependent models, we used gender-dependent decision trees. After clustering, there were about 9,400 distributions for the male models and about 9,500 distributions for the female models.

We performed an open-vocabulary test on the evaluation data (Nov92-20k-si-nvp) where the out-of-vocabulary (OOV) rate is 1.2%. As shown in Table 1, the word error rate for the baseline system is 11.84% with quintphone clustering. The error rate drops to 11.20% if also considering the beginning and ending of words (word boundaries), where we had separate questions for the beginning and ending of a word rather than just one question for the cross-word. The error rate is further reduced to 10.88% if we take into account the word boundaries and function words in addition to quintphone contexts in the decision trees.

System	Word Error
quintphone	11.84%
+word boundary	11.20%
+word boundary +function word	10.88%

Table 1: The word error rates of using different knowledge sources in the decision trees tested on evaluation data with a 30k vocabulary.

5 CONCLUSION

The clustering (or state tying) was originally motivated by the fact that distributions with less or no training observations could be better modeled and also model size could be reduced. Since the phoneme context is only one of the contexts in the whole acoustic space, we may generalize the clustering capability to use more contexts. In this paper we developed an efficient method to cluster not only phoneme contexts but also high-level knowledge sources. Particularly, simple VQ models were used to reduce the computational cost in building the decision trees. For using high-level knowledge sources, the second pass was designed such that it can provide a general type of information for decision trees without compromising the speed of our system. Experimental results suggest that using high-level knowledge sources may increase the recognition accuracy in speech recognition.

The method we have developed provides a general framework to use high level knowledge sources in clustering allophones. There are many factors that influence human perception of speech sounds. For automatic speech recognition, what kinds of factors are important may probably only be judged by careful experiments. Hopefully an efficient and general algorithm may provide us a tool to search for those factors quickly.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, and M.A. Picheny, "Context-dependent modeling of phones in continuous speech using decision trees," *DARPA Workshop on Speech and Natural Language*, pp. 264–269, February 1991.

[2] S.J. Young, J.J. Odell and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," *ARPA Workshop on Human Language Technology*, pp. 286–291, March 1994.

[3] R. Kuhn, A. Lazarides, Y. Normandin and J. Brousseau, "Improved decision trees for phonetic modelling," *Proceedings ICASSP 95*, vol. 1, pp. 552–555, May 1995.

[4] M.Y. Hwang. F. Alleva and X. Huang, "Senones, multi-pass search and unified stochastic modeling in Sphinx-II," *Proceedings Eurospeech 93*, vol. 3, pp. 2143–2146, September 1993.

[5] V. Digalakis and H. Murveit, "Genones: Optimizing the degree of mixture tying in a large vocabulary hidden Markov model based speech recognizer," *Proceedings ICASSP 94*, vol. 1, pp. 533–536, April 1994.

[6] G. Boulianne and P. Kenny, "Optimal tying of HMM mixture densities using decision trees," *Proceedings ICSLP 96*, vol. 1. pp. 350–353, October 1996.

[7] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev and S.J. Young, "The 1994 HTK large vocabulary speech recognition system," *Proceedings ICASSP 95*, vol. 1, pp. 73–76, May 1995.

[8] Z. Li, G. Boulianne, P. Labute, M. Barszcz, H. Garudadri and P. Kenny, "Bi-directional graph search strategies for speech recognition," *Computer Speech and Language*, vol. 10, pp. 295–321, 1996.

[9] Z. Li, M. Heon and D. O'Shaughnessy, "New developments in the INRS continuous speech recognition system," *Proceedings ICSLP 96*, vol. 1, pp. 2–5, October 1996.

[10] Z. Li, P. Kenny and D. O'Shaughnessy, "Hybrid hidden Markov models in speech recognition," *Proceeding Eurospeech 95*, vol. 1, pp. 795–798, September 1995.