S14.8

ALLOPHONE CLUSTERING FOR CONTINUOUS SPEECH RECOGNITION*

Kai-Fu Lee, Satoru Hayamizu, Hsiao-Wuen Hon Cecil Huang, Jonathan Swartz, Robert Weide**
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

As more variabilities are modeled using phone-like subword units, the number of models increases drastically. This paper discusses two techniques for subword clustering that limit this proliferation of models. The first algorithm is based on information theoretic agglomerative clustering, and the second algorithm is based on decision trees. Our preliminary results show that both techniques lead to high performance recognition. We also show that decision tree-based clustering is more robust for vocabulary-independent recognition.

1. Introduction

With recent interest in large-vocabulary systems, subword modeling has become a very important research area. Researchers have found that detailed subword modeling techniques such as context-dependent phone models [1, 2, 3] led to very good results. However, as more detail is modeled, the number of models increases drastically. Even with increased training data, we cannot hope to adequately train models that account for all permutations of contextual variations. In this paper, we describe two methods to limit the proliferation of models through subword clustering.

We define an *allophone* to be a phone in a particular environment [4]. This paper addresses the modeling of environemental variations caused by context: left phoneme, right phoneme, syllable boundary, word boundary, stress of current, previous, and next syllable, and utterance position. Both of our approaches first train rough allophone models, and then cluster them into *generalized allophones*.

The first subword clustering algorithm is based on agglomerative clustering [5, 6], using an information theoretic distance metric. The agglomerative algorithm is excellent from the viewpoint of minimizing entropy, but it has two drawbacks: (1) if an unit requires smoothing, only the context-independent phone is available, which is not sufficiently detailed, and (2) if an allophone in the test data is not covered in the training data, a context-independent phone must be used, which will degrade accuracy [7].

The second technique addresses these problems by using a divisive subword clustering algorithm, based on decision trees [8]. Here the allophones are recursively split by asking questions about contexts (such as *is the next phoneme a back vowel?*.). These questions are first created using human speech knowledge, and the tree is automatically constructed by searching for simple as well as composite questions. The leaf nodes of the tree represent the *generalized allophones* to be used. This tree structure enables smoothing with all ancestor nodes, and context-dependent internal models can be used even if an allophone was never observed before.

We have implemented these algorithms for the DARPA Resource Management task. Preliminary results show that when training data is available for the task, both clustering techniques give comparable results; however, when training on one vocabulary and testing on the other, the decision tree attains slightly better results.

In this paper, we first discuss allophonic modeling in Section 2. In the next two sections, we present the agglomerative and the tree-based clustering algorithms in the first two sections. In Section 5, we describe our experimental set-up, and discuss our results. Finally, some conclusions are given in Section 6.

2. Allophonic Models

Phonemes are natural units of speech; however, because phonemes are highly affected by context, phoneme models have very broad distributions, and are not adequate for high-performance speech recognition. In view of this, triphones [1] were introduced to model every phone in the context of its left and right neighbors. This models the two most important sources of phonetic variability. Triphone-based systems [9, 6] have produced very good results for a 1000-word vocabulary.

However, as we aim for larger vocabulary and higher accuracy, we must consider other sources of variabilities that affect the realization of a phoneme. Figure 1 illustrates some of these sources of variabilities. Triphones are but two of many variabilities that affect the realization of a phone.

In order to minimize the variance within a speech unit, we must explicitly consider each of these sources of phonetic variability. However, the number of combinations is astronomical, which makes training impossible. To limit the proliferation of models, we first use our speech knowledge to identify and model only the most relevant contexts (such as

*This research was sponsored in part by US WEST and in part by the Defense Advanced Research Projects Agency (DOD), Arpa Order No. 5167, under contract number N00039-85-C-0163.

**Visiting Scientist from Electrotechnical Laboratory, Japan

CH2847-2/90/0000-0749 $1.00 © 1990 IEEE

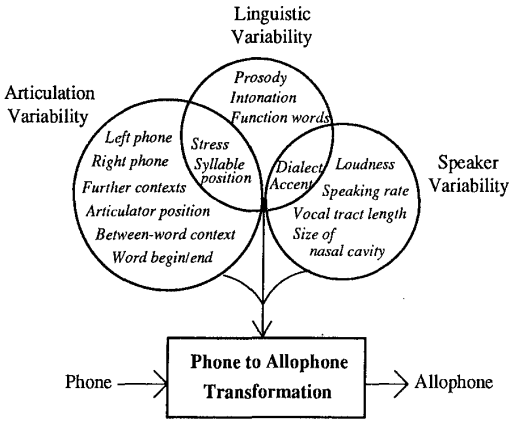

Figure 1: Sources of variability that affect the realization of a phone.

immediate phonetic context, word boundary, and stress). This strategy allows us to reduce an astronomical number of models into a more reasonable number (about 17,000).

Although we have a very large training database, 17,000 models are still too many to train. We use subword clustering to further reduce these models by combining similar ones. This is phonetically plausible because many triphones are, in fact, similar (For example, /b/ and /p/ have similar effects on neighboring vowels). In the next two sections, we describe two methods of subword clustering that reduce 17,000 allophones to about 2,000 generalized allophones.

3. Agglomerative Clustering

Agglomerative clustering of subword units has been used successfully [10, 2]. Our algorithm has been described in [2], and is repeated below:

1. Generate an HMM for every allophone context.
2. Create clusters of allophones; initially, each clusters consists of one allophone.
3. Find the *most similar* pair of clusters which represent the same phone, and merge them.
4. For each pair of same-phone clusters, consider moving every element from one to the other.
    1. Move the element if the resulting configuration is an improvement.
    2. Repeat until no such moves are left.
5. Until some convergence criterion is met, go to step 2.

To determine the distance between two models, we use the following distance metric:

$$D(a,b) = P(m)H(m) - P(a)H(a) - P(b)H(b)$$

$$H(x) = -\sum_{c}^{C} P(c|x) \log P(c|x)$$

where $D(a,b)$ is the distance between two models of the same phone in context a and b, $H(x)$ is the entropy of the distribution in model x, $P(x)$ is the frequency (or count) of a model, and $P(c|x)$ is the output probability of codeword c in model x, In measuring the distance between the two models, we only consider the output probabilities, and ignore the transition probabilities, which are of secondary importance. This information-theoretic distance measure has been shown to be equivalent to a maximum likelihood metric [11].

Although the data-driven agglomerative clustering is a good technique for minimizing entropy, it has two drawbacks. First, a generalized allophone can only be smoothed with the monophone. This smoothing is too crude, especially when the generalized allophone is not well-trained. The other drawback is under vocabulary-independent conditions, when an allophone in the test set has not been observed in the training, the monophone model must be used for the unobserved allophones. This will severely degrade performance [7].

4. Decision Tree Clustering

In order to address these problems, we use decision trees [8, 4, 12] to cluster subword models. At the root of the decision tree is the set of all allophones corresponding to a phone. Each node has a binary "question" about some context of the allophones, i.e., *"is the previous phone a front vowel?"* These questions are generated by an expert linguist, and are designed to capture classes of contextual effects. To find the generalized allophone for an allophone, the tree is traversed by answering the questions attached to each node, until a leaf node is reached. An example of a decision tree for the phone /k/, along with some actual questions, are shown in Figure 2.

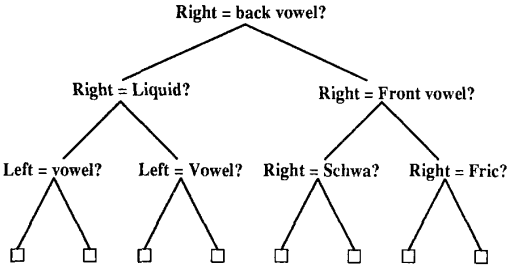

Figure 2: An example of a decision tree that clusters the allophones of the phone /k/

To generate a decision tree for a phone, an HMM is trained for each allophone. All allophones are placed in the root of the initial tree. The "best" question is then chosen from a list of categorical questions to split the "best" node into two nodes. The metric for splitting identical to that in the agglomerative method. We want to find the question that divides node m into nodes a and b, such that

$P(m)H(m) - P(a)H(a) - P(b)H(b)$ is maximized.

This algorithm is then iterated, splitting the most promising node with the most promising split at each iteration. Termination is often best determined by cross-validation with an independent set [8].

One problem with the above algorithm is that with simple questions, the data may be over-fragmented, resulting in similar leaves in different locations of the tree. We deal with this problem by using composite questions [12], or questions that involve conjunctive and disjunctive combinations of all questions (and their negations). A good composite question is formed by first growing a tree using simple questions only, and then clustering the leaves into two sets. Figure 3 shows the formation of one composite question.

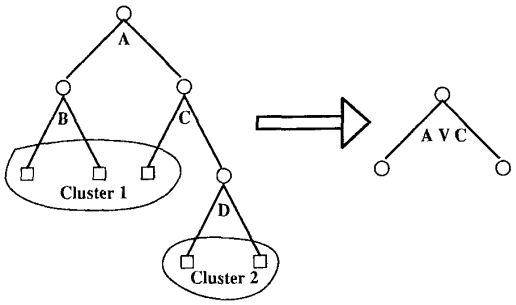

Figure 3: The use of simple-question clustering to form a composite question.

In summary, the decision tree-based algorithm is given below:

1. Generate an HMM for every allophone.
2. Create a tree with one (root) node, consisting of all allophones.
3. Find the best composite question for each node.
    1. Generate a tree with simple questions at each node.
    2. Cluster leaf nodes into two classes, representing the composite question.
4. Split the node with the overall best question.
5. Until some convergence criterion is met, go to step 3.

Tree-based subword clustering overcomes the two problems of agglomerative methods. For smoothing, since a node is similar to its parent node, we could use deleted interpolation [13] to smooth each node with all of its ancestors. Moreover, if a new allophone is encountered, we can traverse the decision tree as deeply as possible, and use the final node as the context-dependent model.

One disadvantage of the tree-based approach is that its clusters are constrained by the questions, so that it is not as powerful in maximizing the objective function as the agglomerative method. However, with a better smoothing method, perhaps more models could be trained, which may compensate for this problem. Another alternative is to use a hybrid approach where decision trees are grown for a small number of steps, and then the leaves are clustered using the agglomerative clustering. This is, in fact, the approach we took in this paper.

5. Experiments and Results

We evaluated these techniques on the speaker-independent DARPA Resource Management database [14]. This task is a 991-word continuous speech task. The word pair grammar (perplexity 60) was used throughout. The test set consists of 320 sentences from 32 speakers (a random selection from the 1988 and 1989 test sets).

For this study, we limit the types of questions to: (1) Left and right phonetic class, (2) stress, and (3) word boundary. A total of 176 questions were generated. Left and right contexts were limited to within-word context. Between-word modeling for allophones is currently being implemented. An older version of the Sphinx System [6] is used for training and recognition.

In the first experiment, we trained the system with a total of 3990 sentences from 109 speakers. For agglomerative clustering, we generated 1000 models, which was shown [2] to be optimal for this task and training size. For decision-tree clustering, we generated phonetic trees (one for each phone) with a total of about 300 nodes. Each node is further divided using the agglomerative algorithm, resulting in 1200 models. More models were used here because we believe the tree-based algorithm can support more models through superior smoothing capability. Table 1 shows the recognition results. Both approaches attain comparable results - about an 8% error rate. Note that between-word modeling and corrective training were not used, so that the error rates are 30-50% higher than our current best system.

Error Rate - RM Training	
Agglomerative Clustering	**Tree-Based Clustering**
8.5%	8.3%

Table 1: Results using vocabulary-dependent training.

In the second experiment, we evaluated these two techniques when some allophones in the test set are missing from the training. We trained the system with a total of 15,000 General English sentences. 5,000 of these were the TIMIT and Harvard sentences [7], and 10,000 were collected at Carnegie Mellon. Here, agglomerative clustering led to about 2,000 models. The decision-tree clustering used a tree with about 400 leaf nodes, which are further divided using agglomerative clustering into about 3,000 models. The 15,000 training sentences cover about 90% of the allophones in the test set. For agglomerative clustering, when an allophone is not trained, the corresponding monophone is used; for decision-tree clustering, the tree is traversed, and the final node in the traversal is used. Table 2 shows the vocabulary-independent recognition results. These results show that with less than full coverage of the test set allophones, the top-down method gives superior performance. We expect that this gap will grow as training data and allophone coverage is further reduced.

Error Rate - GE Training	
Agglomerative Clustering	**Tree-Based Clustering**
15.8%	15.0%

Table 2: Results using vocabulary-independent training

The agglomerative clustering algorithm has already been used successfully. These results show that decision tree clustering is equally powerful, particularly for vocabulary-independent situations. The relatively large gap between vocabulary-dependent and vocabulary-independent conditions was surprising to us, in view of the encouraging results reported in [15]. We conjecture that it may be due to several reasons: (1) the vocabulary-independent training is only four times that of vocabulary-dependent training, while the corresponding ratio in [15] is 15, (2) the triphone coverage is only 90% for this study, and is over 99% for [15], and (3) the recording conditions for training and test material are more different in this study.

In any case, these results indicate that much work remains to be done in the area of vocabulary-independent recognition. We believe that decision tree clustering has much potential that remains to be explored.

6. Conclusion

In this paper, we have presented two methods for subword clustering. The first method is an agglomerative clustering algorithm. This method is completely data-driven, and finds clusters without any external guidance. The second method uses *decision trees* for clustering. This method uses an expert-generated list of questions about contexts, and recursively selects the most appropriate question to split the allophones.

Our preliminary results showed that when the training set has a good coverage of the allophonic variations in the test set, both methods are capable of high-performance recognition. However, under vocabulary-independent conditions, the tree-based allophones outperformed agglomerative clustering because of its superior generalization capability.

Our future work involves more complete implementations and evaluations of both techniques under vocabulary-independent conditions. We also plan to incorporate additional sources of phonetic variability (between-word, syllable boundary, between-speaker variations, etc.). Some of these results will be presented at the conference.

References

1. Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., Makhoul, J., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1985.
2. Lee. K.F., Hon, H.W., Hwang, M.Y., Mahajan, S., Reddy, R., "The SPHINX Speech Recognition System", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1989.
3. Bahl. L.R., et. al, "Large Vocabulary Natural Language Continuous Speech Recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1989.
4. Sagayama, S., "Phoneme Environment Clustering for Speech Recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1989.
5. Duda, R. O., Hart, P. E., *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, N.Y., 1973.
6. Lee, K.F., *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, 1989.
7. Hon, H.W., Lee, K.F., "On Vocabulary-Independent Speech Modeling", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1990.
8. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., *Classification and Regression Trees*, Wadsworth, Inc., Belmont, CA., 1984.
9. Chow, Y.L., Dunham, M.O., Kimball, O.A., Krasner, M.A., Kubala, G.F., Makhoul, J., Roucos, S., Schwartz, R.M., "BYBLOS: The BBN Continuous Speech Recognition System", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1987, pp. 89-92.
10. Hayamizu, S., Tanaka, K., Ohta, K., "A Large Vocabulary Word Recognition System Using Rule-Based Network Representation of Acoustic Characteristic Variations", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1988, pp. 211-214.
11. Lee, K.F., "Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, April 1990.
12. Bahl, L.R., Brown, P.F., de Souze, P.V., Mercer, R.L., "A Tree-Based Statistical Language Model for Natural Language Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-37, No. 7, July 1989, pp. 1001-1008.
13. Jelinek, F., Mercer, R.L., "Interpolated Estimation of Markov Source Parameters from Sparse Data", in *Pattern Recognition in Practice*, E.S. Gelsema and L.N. Kanal, ed., North-Holland Publishing Company, Amsterdam, the Netherlands, 1980, pp. 381-397.
14. Price, P.J., Fisher, W., Bernstein, J., Pallett, D., "A Database for Continuous Speech Recognition in a 1000-Word Domain", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1988.
15. Hon, H.W., Lee, K.F., Weide, R., "Towards Speech Recognition Without Vocabulary-Specific Training", *Proceedings of Eurospeech*, September 1989.