Recent Progress in Robust Vocabulary-Independent Speech Recognition

Hsiao-Wuen Hon and Kai-Fu Lee

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

This paper reports recent efforts to improve the performance of CMU's robust vocabulary-independent (VI) speech recognition systems on the DARPA speaker-independent resource management task. The improvements are evaluated on 320 sentences that randomly selected from the DARPA June 88, February 89 and October 89 test sets. Our first improvement involves more detailed acoustic modeling. We incorporated more dynamic features computed from the LPC cepstra and reduced error by 15% over the baseline system. Our second improvement comes from a larger training database. With more training data, our third improvement comes from a more detailed subword modeling. We incorporated the word boundary context into our VI subword modeling and it resulted in a 30% error reduction. Finally, we used decision-tree allophone clustering to find more suitable models for the subword units not covered in the training set and further reduced error by 17%. All the techniques combined reduced the VI error rate on the resource management task from 11.1% to 5.4% (and from 15.4% to 7.4% when training and testing were under different recording environment). This vocabulary-independent performance has exceeded our vocabulary-dependent performance.

Introduction

As speech recognition flourishes and new applications emerge, the demand for vocabulary-specific training will become the bottleneck in building speech recognizers. If successful, a vocabulary-independent (VI) speech recognition system trained on a large database alleviates the tedious vocabulary-specific training process. We have previously demonstrated the feasibility of vocabulary-independent speech recognition systems [4, 5]. Although the vocabulary-independent results improved as the training data increases, the best vocabulary-independent result previously reported was still about 30% worse than the vocabulary-dependent (VD) result. In this paper, we will report recent efforts to further improve CMU's robust vocabulary-independent speech recognition systems on the DARPA speaker-independent resource management task.

Our first improvement involves the incorporation of more dynamic features in the acoustic front-end processing [7]. Our previous vocabulary-independent experiments have used only first order differenced cepstra and power. Here, we add second order differenced cepstra and power. We also incorporate both 40 msec and 80 msec differenced cepstra. These new features yielded a 15% error rate reduction, about the same as was achieved on vocabulary-dependent tasks [7].

Our second improvement involves the collection of more general English data, from which we can model more phonetic variabilities, such as the word boundary context. Our experiment shows that adding 5,000 sentences to an original 15,000 sentence training set gives only a 3% error reduction. In this experiment, the set of models was fixed.

Next, we incorporated word boundary context into our VI subword modeling, which resulted in a surprising 30% error reduction. This compares with only a 20% error reduction obtained in the vocabulary-dependent case. In the past, it has been argued that between-word triphones may be learning grammatical constraints instead of modeling acoustic variations. This result shows the contrary, since in vocabulary-independent experiments, grammars in the training and recognition are completely different.

With more detailed models (such as between-word triphones), coverage on new tasks was reduced. To deal with this problem, we proposed a new decision-tree based subword clustering algorithm to find more suitable models for the subword units not covered in the training set [11]. These questions were first created using human speech knowledge, and the tree was automatically constructed by searching for simple as well as composite questions. Finally, the tree was pruned using cross validation. When the algorithm terminated, the leaf nodes of the tree represented the *generalized allophones* to be used. This tree structure not only could find suitable models for subword units never observed before, but it also enables smoothing with all ancestor nodes instead of only the context-independent one. In a preliminary experiment, we found that decision-tree based allophones made 17% fewer errors than generalized triphones.

We have found that different recording environments between training and testing (CMU vs. TI) degrades the performance significantly [4], even when the same microphone is used in each case. In [4], we found the vocabulary-independent system suffered much more from differences in the recording environments at TI versus CMU than the vocabulary-dependent system. However, with the above techniques, the vocabulary-independent system became more robust to the changes in recording environment than the vocabulary-dependent system. Our results now show the vocabulary-independent system works about 11% better than vocabulary-dependent system under cross-condition recognition.

These techniques implemented on the vocabulary-independent system led to more than 50% error reduction on both same recording and cross recording conditions. They made our vocabulary-independent system 13% better than our vocabulary-dependent system on the resource management task.

In this paper, we will first describe our recent efforts on CMU's vocabulary- independent speech recognition system, including incorporating more dynamic acoustic feature, larger training database, word boundary context, and decision tree allophone clustering. Then we will describe our experiment setup and present results. Finally, we will close with some concluding remark about this work and future work.

More Detailed Acoustic Model

Temporal changes in the spectra are believed to play an important role in human perception. One way to capture this information is to use differenced coefficients [3, 12] which measure the change of coefficients over time. Our previous vocabulary-independent experiments have used only three codebooks, the first codebook for cepstrum coefficients, the second codebook for differenced cepstrum coefficients (40 msec) and the third codebook for power and differenced power (40 msec) [4].

In additional to first order differencing, it has recently been shown that adding second order differenced coefficients further enhances performance [7]. Thus, we added the fourth codebook with second order cepstrum coefficients. We also incorporated both 40 msec and 80 msec differenced cepstrum coefficients into the second codebook and power, differenced power (40 msec), and second order differenced power into the third codebook. (For detailed implementation, see [7]). These new features reduced the error rate on the vocabulary-independent system from 11.1% to 9.4% and therefore yielded a 15% error reduction, about the same as was achieved on the vocabulary-dependent system [6].

Larger Training Database

In previous work [4], we showed the vocabulary-independent results improved dramatically as the vocabulary-independent training increased. (The error rate was reduced 45% when VI training database was increased from 5,000 sentences to 15,000 sentences). Therefore, we continue collecting more general English database and hope to improve VI results from more VI training.

In addition to the TIMIT (3,300 sentences), Harvard (19,00 sentences) and old general English (10,000 sentences) databases used in the previous experiments, we add 5,000 more general English data into our vocabulary-independent training set. The database covered about 22,000 different words and 13,000 different triphones (not counting inter-word triphones). While the word coverage on DARPA resource management task was only improved from 57% to 60%, the intra-word triphone coverage was improved from 90.0% to 93.6%. We first clustered the 13,000 different triphone models down to about 2,200 by using an agglomerative clustering algorithm [10] and trained on those 2,200 generalized triphone models. However, we only obtained a small improvement, reducing the error rate from 9.4% to 9.1% (a 3% error reduction), when the training database increased from 15,000 sentences to 20,000 sentences.

We conjectured the current subword modeling technique may have reached an asymptote, so that additional sentences are not giving much improvement. If this is correct, we need to make our subword models more detailed with the growing database.

Between-Word Triphone

Because our subword models are phonetic models, one way to model more acoustic-phonetic detail is to incorporate more context information, e.g. stress, word-boundary context, syllable position, etc. We have already incorporated the stress into vocabulary-dependent and vocabulary-independent systems and did not get any improvement [4, 11]. It might be because lexical stress does not predict sentential stress well.

As suggested by the incorporation of word boundary context into triphone modeling in vocabulary-dependent systems [9, 14], we decided to do between-word triphone modeling on our vocabulary-independent system by adding three more contexts, word beginning, word ending and single-phone word positions. The incorporation of word boundary context increased the number of triphones on the VI training set from 13,000 to 33,500 and reduced the triphone coverage on resource management task from 93.6% to 90.0%. We used the same clustering algorithm to cluster those 33,500 triphones down to 2,600 generalized triphones. The between-word triphone modeling enables us to reduce the error rate of the vocabulary-independent system from 9.1% to 6.5%, which is about 29% error reduction.

This result is surprising when compared to only a 20% error reduction in the vocabulary-dependent system [8]. In the past, it has been argued that between-word triphones might be learning grammatical constraints instead of modeling acoustic-phonetic variations. This result shows the contrary, since in vocabulary-independent systems, grammars in the training and recognition are completely different.

Decision Tree Allophone Clustering

As indicated in the previous section, with more detailed models (between-word triphone models), coverage on new task was reduced. For example, the triphone coverage on resource management was reduced from 93.6% (only intra-word triphones) to 90.0% (incorporated inter-word triphones). This means for 10% of the phones in the dictionary, we couldn't find suitable generalized triphone models, and were forced to used the monophone model. This could hurt the system's performance.

To deal with this problem, we proposed a new decision tree based subword clustering algorithm [11, 2, 1, 15]. At the root of the decision tree is the set of all triphones corresponding to a phone. Each node has a binary "question" about their contexts including left, right and word boundary contexts (e.g., *"Is the right phoneme a back vowel?"*). These questions are created using human speech knowledge and are designed to capture classes of contextual effects. To find the generalized triphone for a triphone, the tree is traversed by answering the questions attached to each node, until a leaf node is reached. Figure 1 is an example of a decision tree for the phone /k/, along with some actual questions.

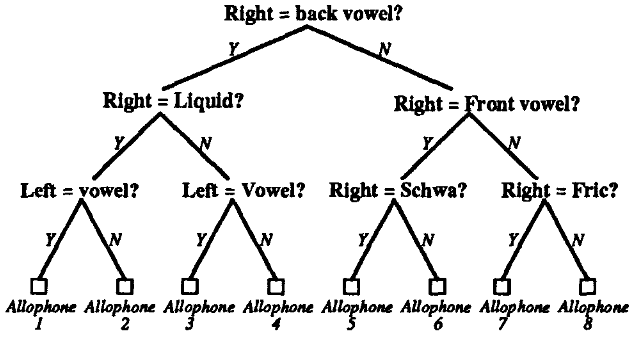

Figure 1: An example of a decision tree that clusters the allophones of the phone /k/

The metric for splitting is a information-theoretic distance measure based on the amount of entropy reduction when splitting a node. We want to find the question that divides node m into nodes a and b, such that

$$P(m)\,H(m) - P(a)\,H(a) - P(b)\,H(b) \text{ is maximized}$$

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

where $H(x)$ is the entropy of the distribution in HMM 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. The algorithm to generate a decision tree for a phone is given below [2]:

1. Generate an HMM for every triphone.
2. Create a tree with one (root) node, consisting of all triphones.
3. Find the best composite question for each node.
   (a) Generate a tree with simple questions at each node.
   (b) 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.

If only simple questions are allowed in the algorithm, the data may be over-fragmented, resulting in similar leaves in different locations of the tree. Therefore, We deal with this problem by using composite questions [1, 13] (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 2 shows the formation of one composite question.

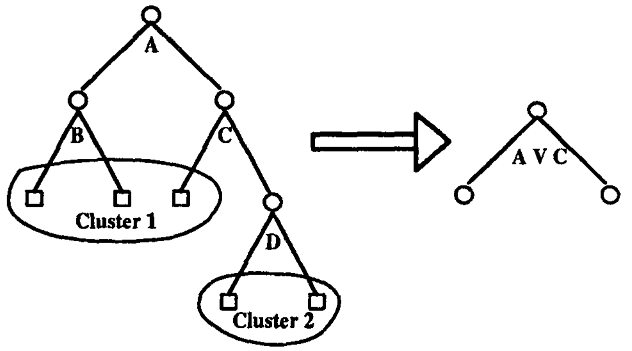

Figure 2: The use of simple-question clustering to form a composite question

To enhance the ability the decision tree clustering to predict the suitable classes for new triphones, we grew the tree a little further and pruned the tree by cross-validation with an independent set [2]. Two thirds of the VI training data was used to train the triphone models and the models were then used to grow the trees. Finally, the other set of triphone models trained from the remaining one third of training data was used to prune the trees.

The tree structure not only could find suitable models for subword units never observed before, but also enable smoothing with all ancestor nodes instead of only context-independent one in the traditional generalized triphone scheme which used agglomerative clustering algorithm. Thus we expected the decision tree based clustering would perform better than other algorithms in vocabulary-independent systems. In a preliminary experiment, the use of decision tree based generalized triphones rather than traditional generalized triphones reduced the error rate of the VI system from 6.5% to 5.4% (a 17% error reduction).

In [4], we found that decision tree based clustering worked only marginally better than agglomerative clustering. The significant improvement here is due to three reasons: (1) improved tree growing and pruning techniques, and (2) our models in this study are more detailed and consistent, which makes it easier to find appropriate and meaningful questions, and (3) triphone coverage is lower in this study, so decision tree based clustering is able to find more suitable models.

Experiments and Results

All the experiments are evaluated on the speaker-independent DARPA resource management task. This task is a 991-word continuous speech task and a standard word-pair grammar with perplexity 60 was used throughout. The test set consists of 320 sentences from 32 speakers (a random selection from June 1988, February 1989 and October 1990 DARPA evaluation sets)

For the vocabulary-dependent (VD) system, we used the the standard DARPA speaker-independent database which consisted of 3,990 sentences from 109 speakers to train the system under different configurations. The baseline vocabulary-independent (VI) system was trained from a total of 15,000 VI sentences. 5,000 of these were the TIMIT and Harvard sentences and 10,000 were General English sentences recorded at CMU. We have shown that different recording environments between training and testing degrades the performance significantly [4]. While the VD training set were recorded at TI, the VI training set were recorded at CMU. Therefore, we recorded at CMU another exactly test set from 32 speakers (different from TI speakers), each speaking 10 sentences (same as the TI sentences), to illustrate the influence of different recording environments. From now on, we use "CMU test set" to denote the test set recorded at CMU and "TI test set" to denote the test set recorded at TI.

In the baseline systems, both the VD and VI systems only used 3 codebook and intra-word generalized triphones. In the first experiment with more acoustic dynamic features, both VD and VI used 4 codebook configuration and got roughly the same improvements. After that, we added 5,000 more general English sentences to the VI training set. We then incorporated inter-word triphones into both VD and VI systems. The VI system was improved more than the VD system. Finally, we used decision tree based generalized triphones on both VD and VI systems. As we expected, the decision tree clustering further improved the VI system by finding more suitable models for subword units never observed in VI training set. Decision tree clustering did not improve the VD system since all the triphones were covered in the VD system. Table 1 shows the recognition error rate for these experiments when training and testing are under the same recording environments. The VI systems was tested on CMU test set and the VD systems was tested on TI test set. Note that the last column showed the percentage of the increase of error rate for the VI system in comparison with the VD system. With the above techniques, the final VI system was better than the VD system.

Configuration	VD	VI	Increase in Error Rate
Baseline	8.6%	11.1%	+29.0%
+ 4 codebooks	7.5%	9.4%	+25.3%
+ 5,000 sentences (VI)	7.5%	9.1%	+21.3%
+ inter-word triphones	6.0%	6.5%	+8.3%
+ decision-tree clustering	6.2%	5.4%	−12.9%

Table 1: The VD and VI results under the same recording condition

The recording environment difference is unavoidable in the vocabulary-independent speech recognition system because the system is trained only once, and must be applied to any applications which could take place in other different environments. In [4], we found the VI system suffered much more from cross environment recognition than the VD system. Table 2 showed the cross environment recognition for both the VD and VI systems. That is, the VI systems were tested on TI test set and the VD systems were tested on CMU test set. We find that the VI system became more robust to the changes in recording environment than the VD system when the VI system had more training data and better subword models. At last, the VI system also performed better than the VD system under cross recording condition.

Configuration	VD	VI	Increase in Error Rate
Baseline	10.8%	15.4%	+42.6%
+ 4 codebooks	10.1%	13.7%	+35.6%
+ 5,000 sentences (VI)	10.1%	12.7%	+25.7%
+ inter-word triphones	8.1%	8.0%	−1.2%
+ decision-tree clustering	8.3%	7.4%	−10.8%

Table 2: The VD and VI results under the cross recording condition

Finally, we tested the last two systems(incoporating inter-word triphones and decision-tree clustering) on the no-grammar recognition. Table 3 showed both results under the same or cross recording conditions. Like the recognition with word-pair grammar, the use of decision tree clustering algorithm reduced the error rate of VI system from 27.8% to 22.8%(a 18% error reduction) under the same recording condition and also made the best VI system better than the best VD system under the same and cross recording conditions.

Configuration	VD	VI	Increase in Error Rate
w/o decision-tree (same)	24.5%	27.8%	+13.5%
w/o decision-tree (cross)	29.2%	30.8%	+5.5%
w decision-tree (same)	25.2%	22.8%	−9.5%
w decision-tree (cross)	29.9%	28.1%	−6.0%

Table 3: The VD and VI results for no-grammar recognition

Conclusions

In this paper, we have presented several techniques that substantially improve the performance of CMU's vocabulary-independent speech recognition system. These techniques, including more dynamic features in acoustic modeling, more training data, more detailed subword modeling (incorporating the word boundary contexts) and decision tree allophone clustering, led to more than 50% error reduction on both same recording and cross recording conditions. This also made our vocabulary-independent system better than our vocabulary-dependent system on the resource management task under both conditions.

In the future, we expect to further extend some of these areas. We will enhance our subword units by modeling more acoustic-phonetic variations, e.g., contexts further than left and right contexts, and function word contexts, etc. Currently, since the use of composite questions might lead to some unreasonable combinations of simple questions, we would like to refine and constrain the type of questions which can be asked to split the decision tree. We would also like to reduce the training data for the decision tree based generalized allophone system and demonstrate the smoothing power and generalizability of decision tree because it would reduce the coverage of the vocabulary-independent systems for new tasks.

Although the vocabulary-independent recognition results on cross recording condition were improved a lot when we had more training data and better subword modeling, there is still a non-negligible degradation for cross recording condition. In the future, we will implement some environmental normalization techniques to further improve the performance of cross environment conditions. Moreover, we would also like to implement some rapid and non-intrusive task adaptation to make the vocabulary-independent system tailored to the individual task.

To make the speech recognition system more robust for new vocabularies and new environments are essential to make the speech recognition application feasible. Our results have shown that plentiful training data, careful subword modeling, and decision tree based clustering have compensated for the lack of vocabulary and environment specific training. We hope with the additional help of environmental normalization and non-intrusive task adaptation, the vocabulary-independent system can be tailored to any task quickly and cheaply.

Acknowledgements

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. The authors would like to express their gratitude to Professor Raj Reddy and CMU speech research group for their support.

References

[1] 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.

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

[3] Furui, S. *Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum.* **IEEE Transactions on Acoustics, Speech, and Signal Processing**, vol. ASSP-34 (1986), pp. 52–59.

[4] Hon, H. and Lee, K. *On Vocabulary-Independent Speech Modeling.* in: **ICASSP**. 1990.

[5] Hon, H., Lee, K., and Weide, R. *Towards Speech Recognition Without Vocabulary-Specific Training.* in: **Proceedings of Eurospeech**. 1989.

[6] Huang, X., Alleva, F., Hayamizu, S., Hon, H., and Lee, K. *Improved Hidden Markov Modeling for Speaker-Independent Continuous Speech Recognition.* in: **DARPA Speech and Language Workshop**. Morgan Kaufmann Publishers, San Mateo, CA, 1990.

[7] Huang, X., Lee, K., Hon, H., and Hwang, M. *Improved Acoustic Modeling with the SPHINX Speech Recognition System.* in: **ICASSP**. 1991.

[8] Hwang, M. *Personal Communication.* unpublished, 1988.

[9] Hwang, M., Hon, H., and Lee, K. *Modeling Between-Word Coarticulation in Continuous Speech Recognition.* in: **Proceedings of Eurospeech.** 1989.

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

[11] Lee, K., Hayamizu, S., Hon, H., Huang, C., Swartz, J., and Weide, R. *Allophone Clustering for Continuous Speech Recognition.* in: **ICASSP.** 1990.

[12] Lee, K., Hon, H., and Reddy, R. *An Overview of the SPHINX Speech Recognition System.* **IEEE Transactions on Acoustics, Speech, and Signal Processing,** January 1990.

[13] L.R., B. and et. al. *Large Vocabulary Natural Language Continuous Speech Recognition.* in: **ICASSP.** 1989.

[14] Pieraccini, R., Lee, C., Giachin, E., and Rabiner, L. *Implementation Aspects of Large Vocabulary Recognition Based on Intraword and Interword Phonetic Units.* in: **DARPA Speech and Language Workshop.** 1990.

[15] Sagayama, S. *Phoneme Environment Clustering for Speech Recognition.* in: **ICASSP.** 1989.