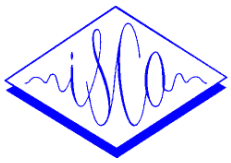

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

DESCRIPTION OF ACOUSTIC VARIATIONS BY TREE-BASED PHONE MODELING

Satoru Hayamizu † Kai-Fu Lee ‡ and Hsiao-Wuen Hon ‡

†Electrotechnical Laboratory, Tsukuba Science City, Japan.
‡Carnegie Mellon University, Pittsburgh, PA 15213 U.S.A.

abstract

This paper discusses the use of tree-based phone modeling to describe acoustic variations of speech, and its application to speech recognition system. There are many sources of variabilities that affect the realization of a phoneme: phonetic contexts, speakers, stress, speaking rates and so on. Explicit modeling with these sources of variabilities will give more accurate and more detailed phone models, but needs a large amount of speech data for training. Tree-based phone modeling is studied to solve this problem with three case studies : phone models with large VQ codebook sizes, decision tree clustering, and speaker-clustering. They are tested on speaker-independent continuous speech recognition experiments with a 991 word vocabulary. Tree-based phone modeling is shown to produce improvement in all three cases and to provide a good guide to provide trainability and generalizability.

1 Introduction

The purpose of this paper is to study tree-based phone modeling for better description of acoustic characteristic variations of speech.

There are many sources of variabilities that affect the realization of a phoneme: phonetic contexts, speakers, stress, speaking rates and so on. It has been found that context dependent phone modeling [21],[6],[13],[2] and clustering of context dependent phones produce very good results in speech recognition. In a more general form, an allophone can be defined as a phone in a particular environment of these sources of variabilities [20],[7],[8],[15]. Explicit modeling with these variabilities will give more accurate and more detailed phone models, but an astronomical amount of training data is needed to train all the allophones sufficiently.

In this paper, tree-based phone modeling is proposed to obtain a better description of acoustic characteristic variations. There are some hierarchies in these phone models from well trained but less accurate models (for example, context-independent phone models) to less trained but more accurate ones (for example, context-dependent phone models). Hierarchical tree structures of phone models are introduced in order to enable incremental training from rough models to detailed models and to enable smoothing with internal nodes in the tree of phone models.

Three case studies of tree-based phone modeling are presented; First, it is applied to phone models with large VQ codebook sizes to get better resolution. Second, it is applied to decision-tree based context clustering [15] for prediction of unknown contexts and better smoothing. Third, it is applied to speaker clustering and description of acoustic variations for both context and speaker. All three cases are tested on speaker-independent continuous speech recognition experiments with a 991 vocabulary using an HMM-based speech recognition system, SPHINX. It is shown that tree-based phone modeling produces improvement in all three cases.

2 Phone Models with Large VQ-Size

2.1 Smoothing Models with Different VQ-Sizes

In dicsrete HMMs, the distortion of vector quantization affects the accuracy of acoustic detail. Smaller the distortion leads to more accurate of modeling. The problem is that if we increase the codebook size, we will need more data to train the HMMs.

Tree-based phone modeling , where each node consists of models with different VQ sizes, will provide better smoothing for that purpose. The idea is to use binary tree searched vector quantization [5] to make the codewords in different size of codebooks related. All the codewords in the large size of VQ codebooks are descendents of the codewords in a smaller size of VQ codebooks.

In smoothing, each node is smoothed using a linear combination of all the nodes in the path from the node up to the root node. Also, a special node of uniform distribution is added on the top of root node in order to avoid a zero probability. For example, models of codebook size 4096, X_{4096} can be smoothed with models of codebook-size 1024, X_{1024}, those of codebook size 256, X_{256} and uniform distribution U as:

$$X'_{4096} = \lambda_1 X_{4096} + \lambda_2 X_{1024} + \lambda_3 X_{256} + \lambda_4 U$$

where $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$. Estimation of λ_i is done by *deleted interpolation* [12]. It is to divide the training data into several blocks, use all the blocks except a *deleted* block to estimate λ_i on that block, and average the λ results.

2.2 Experiments and Results

The database used here is the speaker-independent DARPA Resource Management database [Price 88]. The task is a 991-word continuous speech task. The word pair grammar (perplexity 60) was used with no corrective training. The test set (TI-test) consists of 320 sentences from 32 speakers randomly selected from the 1988 and 1989 test sets. The training set consists of 4,358 sentences from 109 speakers.

A cepstrum analysis of order 14 is done and 32 LPC cepstral coefficients are calculated by a recursive equation.

Then bilinear transformation is done resulting in 12 mel-scaled cepstral coefficients. Frame shift is 10 msec and sampling frequency is 16 kHz. Three codebooks of cepstrum, delta-cepstrum, power-and-delta-power are used. Unless otherwise specified, the recognition system is identical to the version of SPHINX described in [13].

The recognition results using the 48 context independent phones are shown in Table 2.1. Codebook sizes of 256 (standard SPHINX), 512, 1,024, 2,048, 4,096 are used. Models were smoothed with smaller codebooks of models. Deleted interpolation was done based on the count ranges in these experiments. Note, a codebook size of 2,048 actually gave a 1.0% improvement in word accuracy (about 5% error reduction). Table 2.2 shows the results where models were smoothed only with a uniform distribution. In the case of a codebook size of 2,048, smoothing with internal nodes gave a 1.0 % improvement in word accuracy, comparing with smoothing without internal nodes.

Table 2.1 Recognition results for codebook size of 256, 512, 1024, 2048, 4096 (models are smoothed using those with different codebook sizes).

size	percent correct (word accuracy)	smoothing
256	83.0% (80.8%)	baseline
512	83.5% (81.0%)	with 256 and uniform
1,024	83.8% (81.3%)	with 256 and uniform
2,048	84.2% (81.7%)	with 256 and uniform
4,096	84.2% (81.6%)	with 256 and uniform
2,048	84.1% (81.8%)	with 256,512,1024,uniform
4,096	84.2% (81.7%)	with 256,1024,uniform

Table 2.2 Recognition results for codebook size of 512, 1024, 2048, 4096 (smooth models with uniform).

size	percent correct (word accuracy)	smoothing
512	83.9% (81.4%)	with uniform
1,024	83.8% (81.2%)	with uniform
2,048	83.3% (80.8%)	with uniform
4,096	83.4% (80.6%)	with uniform

3 Decision Tree Clustering

3.1 Decision-Tree-based Context Clustering

The agglomerative clustering for generalized triphones used in [13] is excellent from the point of view of minimizing entropy. However, it has two drawbacks : generalized triphones can only be smoothed with context-independent phones alone, and it is not possible to find the generalized triphone for a triphone not observed in training.

Decision-tree-based clustering of phonetic contexts will provide a tree structure for better smoothing and prediction. The idea is to use the features of neighbouring phones to guide context clustering so that the acoustic realization of unknown contexts can be analogically predicted using these features. Also, internal nodes of the tree represent classes of allophones, and provide intermediate representations for better smoothing. Decision trees [4] have been used to get statistical language models [3] and to cluster phones into broad classes. Here the technique is applied to context clustering [2], [15].

Figure 3.1 shows an example of decision tree that clusters the contexts of the phone /k/. Each node has a binary "question" about contexts of the allophones, for example,

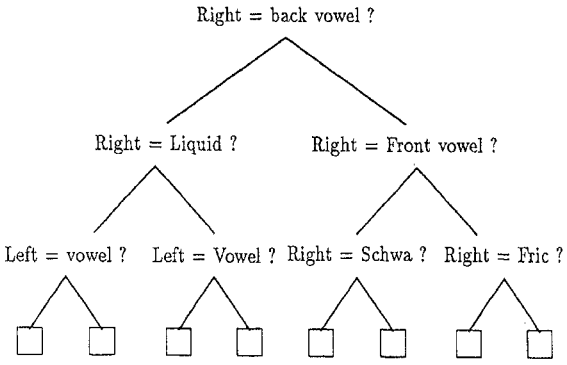

Figure 3.1. An example of a decision tree that clusters the contexts of the phone /k/

"is the previous phone a front vowel ?". Each node represents a sub-set of contexts according to the questions in the path from the root node to itself. The distance metric used for this clustering is identical to that for agglomerative clustering [13]. Other details about decision tree clustering are found in [15].

All the nodes in the tree are smoothed along all the paths from the root node to the node which is to be smoothed. Also, the special node of uniform distribution is added on the top of the root node in order to avoid a zero probability. Estimation of λ_i is done by deleted interpolation.

3.2 Experiments and Results

The task tested here is the same as Section 2.2. A total of 1,800 leaf nodes (models) were generated by the decision tree clustering and agglomerative clustering and used for the recognition experiments. Between-word modeling for allophones and corrective training were not used, thus the error rates are 30-50% higher than the current best version of the SPHINX system.

We tested the decision tree clustering on vocabulary independent recognitions [9], [10]. The General English training set consists of 3,000 TIMIT sentences, 2,000 Harvard sentences, and 10,000 General English sentences which are collected at Carnegie Mellon. The total 15,000 training sentences cover about 90% of the triphones in the test set. The test set consists of a TI-test (same as Section 2.2) and a CMU-test. The CMU-test set consists of 320 sentences (same sentences as TI-test) from 32 speakers (different speakers) recorded at Carnegie Mellon.

Table 3.1 shows the preliminary recognition results for the CMU-test set using the General English training sentences. The error rate of decision-tree clustering is comparable for that of agglomerative clustering [13].

The current triphone coverage of 90% may become larger if it is weighted by frequency and the missing 10% contexts may not be important. So, we tested less covered vocabulary by using the TIMIT training sentences alone.

Table 3.1 Recognition results for the CMU-test set using the General English training sentences

	percent correct (word accuracy)
agglomerative clustering	90.4% (88.9%)
decistion-tree clustering	90.4% (89.2%)

Table 3.2 Recognition results for the TI-test set using the TIMIT training sentences

	percent correct (word accuracy)
agglomerative clustering	84.8% (82.0%)
decistion-tree clustering	85.4% (82.9%)

Table 3.2 shows the recognition results for the TI-test set using the TIMIT training sentences. The word accuracy of decision-tree clustering is 0.9% better than that of agglomerative clustering (5% error reduction). These results indicate decision tree clustering is powerful, particularly for vocabulary independent situations.

However, there is a gap between the error rates for vocabulary dependent and vocabulary independent training sets in the case of the test set recorded at TI [15]. The gap is probably due to the difference in recording conditions between the training and test materials. It indicates that some kind of noise reduction or adaptation to the recording environment [1] is necessary to fill the gap between two training sentences.

4 Speaker Clustering

4.1 Top-Down Speaker Clustering

There are some studies on speaker clustering in a dicrete HMM-based sytem [13] (agglomerative clustering) and in a continuous HMM-based system [19]. In this section, tree-based phone modeling is applied to speaker clustering.

The algorithm of speaker clustering used here is a variant of Linde-Buzo-Gray algorithm [17] for vector quantizer design. It will cluster speakers by top-down splitting [8]. The distance metric of two models is the same as that in [13]. Here, phone models for each speaker are trained using sentences spoken by the speaker. Only output probabilities were considered in the clustering.

Figure 4.1 shows the relationship between the number of speaker clusters and the probabilities that a speaker cluster produces each speaker model with (including the speaker tested) and without (eliminating the speaker tested) cross validation [4]. Probabilities for the generalized triphone of "AE(1)" out of 1,100 generalized triphones of 109 speakers in the Resource Management training sentences are shown. It is shown that probabilities increase constantly without cross validation but those with cross validation decrease after three speaker clusters. And this figure suggests that we can expect some improvement using only two or three speaker clusters.

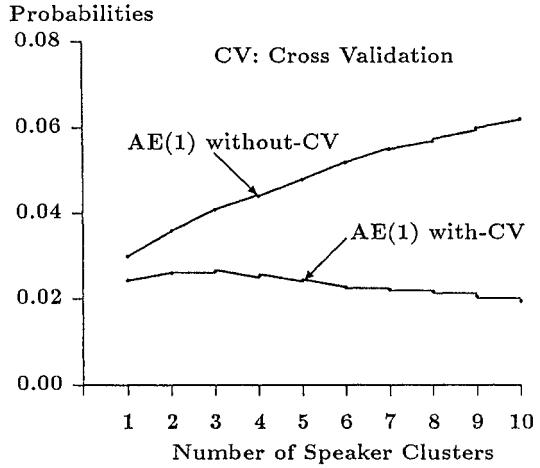

Figure 4.1. Probabilities at Speaker Clustering of Generalized Triphones [AE(1)].

4.2 Smoothing of Speaker-Cluster HMMs

With the combination of contexts and speakers as two sources of variabilites, there are two hierarchies, one is Context-Independent (CI) and Context-Dependent (CD) hierachy and the other is Speaker-Independent (SI) and Speaker-cluster-Dependent (SD) hierarchy.

Acoustic variabilities for speakers are described by multiple speaker clusters. To use these multiple speaker clusters for the recognition, smoothing is necessary to get more robust models. Of the two sources of variabilities, between-context variabilities are believed to be greater than between-speaker ones. So, we simplify the CI-CD, SI-SD hierarchies to the following tree (Figure 4.2).

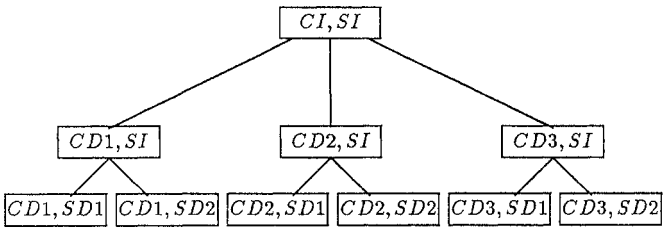

Figure 4.2 A simplified tree of description of variations for contexts $(CI - CD)$ and speakers $(SI - SD)$.

Also, a special node of uniform probabilities is added to the top of root node. Smoothing is conducted by the linear combination of all the nodes from the root node to the node to be smoothed.

$$X'_{CD,SD} = \lambda_1 X_{CD,SD} + \lambda_2 X_{CD,SI} + \lambda_3 X_{CI,SI} + \lambda_4 U$$

where $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$. Estimation of λ_i is done by deleted interpolation. We estimate λ_i for three distributions (begin, middle and end) of each phone along all the paths independently.

4.3 Experiments and Results

First, we conducted a top-down clustering of speakers. We used 4,358 sentences from 109 speakers (about 40 sentences per speaker) to train phone models for each speaker. For the speaker clustering, we used only 47 context-independent phones (silence is excluded) of 109 speakers. We splitted the 109 speakers into two and three speaker clusters. Most clusters are dominated by male or female as the bottom-up speaker clustering [13].

Using these clusters, all speaker clusters were trained using same database of 4,358 sentences from 109 speakers. A total of 1,100 generalized triphones using between-word modeling was used to represent variations for contexts. We used speaker independent models as the initial models and one iteration of forward backward training was run.

The task tested here is the same as Section 2.2. Both cases were tested with word pair grammar (perplexity 60) and without grammar using the TI-test set. All speaker clusters are tested for each speaker, and the one yielding the highest recognition probability (not accuracy) is used. So this speaker-clustering scheme preserves speaker independence. Table 4.1 shows the recognition results using two and three speaker clusters.

Using two speaker clusters, we obtained about 6% (word pair grammar) and 9% (no grammar) error reduction. These results show the potential of speaker clustering. Huang had obtained almost the same results as two speaker cluster case by training male and female separately and by testing them for known gender [11].

Both clustering results and recognition results indicate that we can expect improvement using only two or three speaker clusters. It may be due to several reasons. First, we need a larger database to train each speaker cluster sufficiently. Secondly, we may need to use speaker cluster dependent VQ codebooks, for we used the same VQ codebooks for all the speaker clusters.

Table 4.1 Recognition results using two and three speaker clusters. Results shown are percent-correct (word-accuracy).

Word Pair Grammar	
Speaker Independent	94.2% (93.0%)
2 speaker clusters	94.7% (93.4%)
3 speaker clusters	94.6% (93.4%)
No Grammar	
Speaker Independent	75.3% (72.2%)
2 speaker clusters	77.4% (74.6%)
3 speaker clusters	76.8% (73.8%)

5 Conclusion

In this paper, we have presented a description of the acoustic variations by tree-based phone modeling.

First, phone models with different VQ codebook sizes were studied. Binary tree searched vector quantization was used to make different size of codebooks being related to each other. A codebook size of 2,048 gave about a 5 % error reduction for the case of context independent phones. Smoothing models with different size of codebooks also gave us more robust modelings.

Second, decision tree clustering was presented to provide a tree structure for better smoothing and prediction about unknown contexts. The recognition results were comparable for the General English training set and about a 5% error reduction for the TIMIT database. Decision tree clustering is shown to be powerful, particularly for vocabulary independent situations.

Finally, tree-based phone modeling for speaker clustering was studied. The importance of cross validations for speaker clustering was shown. Using two speaker clusters with 1,100 generalized triphones, we obtained about 6% (word pair grammar) and 9% (no grammar) error reduction and these results show the potential of speaker clustering.

Tree-based phone modeling provides an important step from the current context-dependent phone modeling to a more general description of acoustic characteristic variations. It will provide a good guide to find the compromise of trainability and specificity of phone modeling. But we believe a larger database is still necesssary to get enough training for more detailed description of acoustic variations.

Acknowledgments

The authors would like to thank Professor Raj Reddy for his encouragement and support and would like to thank Mr. Robert Weide for providing the speech database and phonological knowledge for the decision tree clustering. The authors would like to thank Mr. Cecil Huang and Mr. Jonathan Swartz for providing software for decision tree clustering and would also like to thank Miss Jeanette Dravk for revising the CMU technical report.

References

[1] Acero, A., Stern, R.M., "Environmental Robustness in Automatic Speech Recognition", ICASSP-90.
[2] Bahl, L.R., et. al., "Large Vocabulary Natural Language Continuous Speech Recognition", ICASSP-89.
[3] Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L., "A Tree-Based Statistical Language Model for Natural Language Speech Recognition", IEEE Trans. ASSP-37, 7, July 1989.
[4] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., Classification and Regression Trees, Wadsworth, Belmont, 1984.
[5] Gray, R.M., Linde, Y., "Vector Quantizers and Predictive Quantizers for Gauss-Markov Sources", IEEE Trans. COM-30, 2, February 1982.
[6] Hayamizu, S., Tanaka, K., Ohta, K., "A Large Vocabulary Word Recognition System Using Rule-Based Network Representation of Acoustic Characteristic Variations", ICASSP-88.
[7] Hayamizu, S., Tanaka, K., Ohta, K., "On generalized description of acoustic characteristic variations of speech", IEICE of Japan Trans. J72-D-II, 8, August, 1989.
[8] Hayamizu, S., Lee, K.F., Hon, H.W., "Description of Acoustic Variations by Hidden Markov Models with Tree Structure", CMU Technical Report, CMU-CS-90-116, March 1990.
[9] Hon, H.W., Lee, K.F., Weide, R., "Towards Speech Recognition Without Vocabulary-Specific Training", Proceedings of Eurospeech, September 1989.
[10] Hon, H.W., Lee, K.F., "On Vocabulary-Independent Speech Modeling", ICASSP-90.
[11] Huang, X.D., Personal Communication, unpublished, 1990.
[12] 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 Co., Amsterdam, pp,381-397, 1980.
[13] Lee, K.F., Automatic Speech Recognition: The Development of the SPHINX System, Kluwer Academic Publishers, 1989.
[14] Lee, K.F., Hon, H.W., Huang, M.Y., Mahajan, S., Reddy, R., "The SPHINX Speech Recognition System", ICASSP-89.
[15] Lee, K.F., Hayamizu, S., Hon, H.W., Huang, C., Swartz, J., Weide, R., "Allophone Clustering for Continuous Speech Recognition", ICASSP-90.
[16] Lee, K.F., "Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition", IEEE Trans. ASSP-38, 4, April 1990.
[17] Linde, Y., Buzo, A., Gray, R.M., "An Algorithm for Vector Quantizer Design", IEEE Trans. COM-28, 1, January, 1980.
[18] Price, P.J., Fisher, W., Bernstein, J., Pallett, D., "A Database for Continuous Speech Recognition in a 1000-Word Domain", ICASSP-88.
[19] Rabiner, L.R., Lee, C.H., Juang, B.H., Wilpon, J.G., "HMM Clustering for Connected Word Recognition", ICASSP-89.
[20] Sagayama, S., "Phoneme Environment Clustering for Speech Recognition", ICASSP-89.
[21] Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., Makhoul, J., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", ICASSP-85.