ON THE USE OF SPEAKING RATE AS A GENERALIZED FEATURE TO IMPROVE DECISION TREES

R. Faltlhauser, T. Pfau, G. Ruske

Institute for Human-Machine-Communication
Technische Universität München (TUM), Germany
{faltlhauser,pfau,ruske}@ei.tum.de

ABSTRACT

Decision trees are probably the most common way for generating models for phonemes in their phonetic context. In this paper we investigated several ways how speaking rate information can be integrated in the decision tree process. We basically focused on two approaches: on the one hand a speaking rate feature included in the decision tree itself and on the other hand a pruning approach for creating individual model sets. Recently, some papers have come up with the idea to include a gender feature already in the decision process. In our paper we went a step further and wanted to see whether speaking rate can be a fruitful extension to decision trees. Experiments have shown that the introduction of speaking rate leads to improvements in combination with a general gender feature. Further experiments with different pruning strategies aimed at creating adequate model sets for different speaking rate categories.

1. INTRODUCTION

Decision trees are a widespread tool for modelling the phonetic context of a phoneme. Usually this kind of algorithm is used for the tying of triphone (or polyphone) states. In [1] a decision tree is grown for each individual state of the underlying Hidden Markov Model (HMM). The concept of decision trees is rather general and does not have to be restricted to the application of phonetic questions. Paul [2] for example was using a single tree including questions towards phone and state. Wider contexts [3], which are considered for modeling polyphones can be realized by adjusting the question set used for the decisions.

One of the major advantages, that makes the application of decision trees attractive is the fact, that they are able to cover unseen phonetic contexts. The structure of the phonetic decision trees enables mapping triphones not occurring in the training data to a similar model.

As mentioned, a common way to apply the concept of decision trees to the modeling of phonetic context is to use an individual decision tree for each state of the underlying phoneme. Basically the decision tree algorithm is an iterative binary splitting algorithm. Hereby each possible split is determined by a set of phonetic questions. In order to find the optimum phonetic split of a tree node all questions are evaluated based on a quality measure. Usually the likelihood gain ΔL is used:

$$\Delta L = L_{No\_child} + L_{Yes\_child} - L_{parent}$$

whereby the likelihood score L_{node} can be determined from a node model. As node models usually uni- or multimodal Gaussian mixture models are being used.

Mainly because of performance reasons, each node is modeled often only with a single Gaussian density, characterized by the node sample mean and the according variance. The splitting process is continued until some termination criterion is satisfied. Basically the termination is controlled by 2 parameters: a minimum number of feature vectors (samples) per node and a minimum gain threshold ΔL_{min}.

In order to grow trees with an appropriate size several techniques have been proposed. A common approach is the concept of pruning as used by Lazarides et al. [4, 5] or Rogina [6].

2. DECISION TREES WITH GENERALIZED FEATURES

The term "generalized features" in our sense denominates features apart from the usual phonetic context. These features themselves can also be divided in 2 groups: on the one hand static features which - like the phonetic context - do not depend on the actual realization of an utterance. They primarily depend on its phonetic content. A good example for it would be the position of a phoneme inside a word. On the other hand there are dynamic features which describe the way the utterance was spoken, e.g. fast or slow. The more prominent features of this kind are perhaps

- gender and
- speaking rate.

A major problem arising is the question how to incorporate these features in the recognition phase of a speech recognition system. A simple way would be to run a recognizer with each potential category in parallel

From: Proc. ICSLP 2000, IEEE

- deciding afterwards based for example on the best score achieved. This approach however assumes that no change in these dynamic features is occurring during the utterance. Furthermore with an increasing number of generalized features the number of recognizers to be run in parallel is dramatically increasing. Nevertheless this approach offers a first idea whether the use of generalized features offers some improvement at all. Another approach would be to include the information obtained from an estimator [7, 8, 9, 10] directly in the search process.

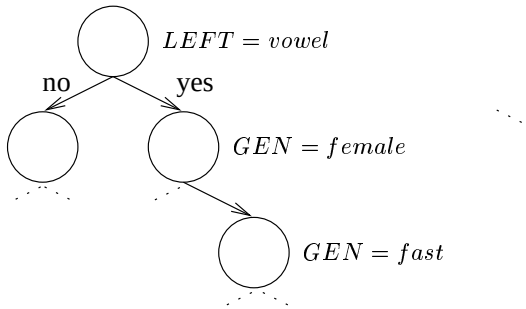

Figure 1: Decision tree including general context.

Figure 1 shows a sample of a decision tree including generalized features. The plot is to illustrate the resulting trees. In contrast to gender, speaking rate is usually not found that near to the root node. A similar approach was recently published by [11].

As explained in [12] the inclusion of general (tagged) features in the decision tree partially circumvents the problem of training data depletion. This problem becomes even more prominent if more generalized features are being used. For example: using the 2 gender categories as well as 3 speaking rate categories would split the whole training data into 6 different bins. Especially for extreme rate categories training data is already rather scarce. Because of the Gaussian-like distribution of speaking rates in the training data, only about 10 percent of all utterances can be considered as fast or slow. Although nearly the same number of utterances are marked as fast and slow, there are far fewer training vectors for fast speech. This results from the shorter duration of phonemes in fast speech. Introducing gender would result in a further division of the already scarce training data.

3. DECISION TREE PRUNING

It has been shown that the training of separate model sets for different speaking rate categories can lead to some improvement. In [13] Bayesian training was used to tackle the problem of data depletion for the extreme rate categories. Using context dependent modeling the question is arising how to generate an appropriate model set for the different categories. The approach we are proposing here is based on a suited tree for each category. Since the growing of an individual decision tree might be difficult with few data we were examining the possibilities of decision tree pruning. The main idea behind this approach is to grow a common base tree for all rate categories which is sligthly overgrown. Afterwards this tree is individually trimmed to suit the different rate categories.

3.1. Pruning with node models

Our first approach was based on the node models used to generate the base tree. As common for the generation of decision trees we were using unimodal Gaussian densities to determine the likelihood of a node split. These densities can be used for crossvalidation on another data set. As crossvalidation data set we used an independent subpart of the training data which was divided into the 3 rate categories slow, average and fast speech. As mentioned, the main idea behind the pruning approach is to grow an overgrown mutual starting tree for all three different rate categories. Beginning with this common tree an individual tree is generated for each rate category by pruning the seed tree using the according crossvalidation data. Pruning is basically carried out by iteratively collapsing two child nodes whose likelihood gain falls below a given threshold. The pruning algorithm starts from the terminal nodes and stops if no further recombination is possible.

3.2. Pruning with recognizer models

As a second pruning approach we were investigating the performance after using full models for pruning. For this purpose we trained a full codebook for each node of the decision tree - not only the terminal nodes. In order to maintain a nearly equal number of Gaussians after the pruning step, we set the codebook size K^c of each node such that:

$$K^c_{parent} \approx K^c_{No\_child} + K^c_{Yes\_child}$$

For the training and segmentation we applied a Segmental K-Means training algorithm. Whereas for the segmentation of the training utterances only the terminal nodes of the tree are used, all inner nodes are updated, too. If a given frame p is assigned to a certain leaf node (triphone state) this node codebook is updated together with all intermediate nodes. The intermediate nodes can be accessed by traversing the decision tree backwards from the terminal nodes to the root node (Figure 2). All tree nodes along this decision path get an update from this particular feature vector. By this way an iterative training of all node codebooks can be achieved.

By aligning the utterances of a crossvalidation set, node statistics can be accumulated. Each frame of an utterance is scored with all intermediate nodes along the root-terminal node path. Now either the score or the node counts are accumulated. The node counts are basically obtained by the evaluation of all nodes along the path and the determination of the node yielding the highest likelihood. The accumulated node scores as well as the node counts allow either an iterative pruning by local decisions

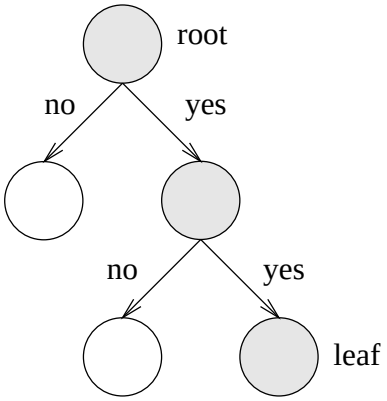

Figure 2: Nodes along the decision path from the root to the leaf node.

starting from the terminal nodes or a pruning of whole subtrees. A local decision is made upon the comparison:

$$L_{parent} - L_{thres} > L_{No\_child} + L_{Yes\_child}$$

If this expression holds true, the two child nodes are collapsed. In our experiments we set $L_{thres} = 0$.

Since the alignment statistics of the whole path is available, long range decisions can also be drawn by determining the optimal values along the different paths.

An advantage of this pruning approach is the fact, that the resulting models after the pruning step could readily be used for recognition. No new initialization is necessary. Nevertheless it emerged, that one training iteration is advantageous to retrain the mixture coefficients.

4. EXPERIMENTAL RESULTS

All experiments were conducted on the German Verbmobil corpus using the eval96 test set. About 11000 utterances from 600 male and female speakers were used for training and the generation of the decision trees. Preprocessing is based on MFCCs with 42 components: 12 MFCCs, total-energy and zero crossing rate as well as first and second derivatives. For the pruning approaches the training data was split into training and crossvalidation data.

4.1. Decision trees with generalized features

In our experiments we defined 3 categories for the speaking rate: slow, average and fast speech. Category boundaries are given by the rate mean ± standard deviation. These labels, together with gender were used for the creation of our decision trees.

To give an impression of the importance of the different general features, Table 1 shows for some phonemes the first occurrence of a question towards gender. The given numbers depict the decision layer in which the decision is drawn. A value of 1 means a decision in the root node.

	first occurrence		
Phoneme	state 0	state 1	state 2
/z/	1	1	1
/s/	2	1	2
/a/	3	2	2
/E:/	2	2	2
/m/	3	3	3
/f/	2	1	4
/t/	3	4	3
/t/	3	3	4

Table 1: First occurrence of question for a gender feature in decision tree.

	first occurrence		
Phoneme	state 0	state 1	state 2
/z/	NA	3	4
/s/	6	4	6
/a/	>6	3	6
/E:/	4	3	4
/m/	3	3	4
/f/	2	4	6
/t/	5	6	6
/t/	NA	4	NA

Table 2: First occurrence of question for speaking rate feature in decision tree.

As can be seen from Table 2 speaking rate is comparatively of 'minor' importance in comparison to gender. Of course, this fact is understandable since extreme rate categories comprise only a small part of the training data. Fast speech for example sums up to only about 10% of the entire training utterances.

Since the focus of our paper is on the use of speaking rate we assumed the gender information to be given. The evaluation utterances were categorized based on the phoneme rate. It was computed on the transliteration by means of a Viterbi alignment.

General features	WER [%]
NONE (baseline)	25.8
SR	27.1
G	24.7
G + SR	24.5

Table 3: WER for trees with different types of generalized context and supervised feature estimation.

As can be seen from Table 3, the introduction of a speaking rate (SR) feature only leads to an overall decrease in recognition performance, although for slow speech an

improvement of 3.5% relatively was observed. This becomes understandable considering the resulting decision trees: most questions concerning speaking rate are questions towards slow speech. In combination with gender an overall improvement can be seen. In correspondence with the models, which only use the speaking rate feature, the highest improvement (7.4% relatively) was achieved for slow speech. Experiments using an estimation of the speaking rate [10] showed similar results.

4.2. Pruning techniques

Pruning with node models

Using a pruning strategy based on simple single Gaussian node models showed improvements up to 5.3% relatively for the slow speech models. The larger baseline system could not be outperformed for fast speech. Nevertheless, the experiments showed that fast speech tends to need more leaf nodes for modeling as slow speech.

Pruning with recognizer models

	pruning data			
rate category	slow	avg.	fast	baseline
all	24.9	24.7	24.8	24.7
very slow	**20.0**	21.5	20.6	22.7
slow	**24.0**	**24.3**	26.9	24.9
average	23.3	**22.6**	26.7	22.6
fast	27.1	**26.0**	**26.7**	27.1
very fast	35.0	34.9	**32.6**	31.4

Table 4: Full models: WER [%] for trees with a different rate category of pruning data each.

Table 4 shows the recognition performance for the baseline system and a specific model set for slow, average and fast speaking rate. All 3 model sets were generated by pruning the decision tree with the according crossvalidation data and retraining the mixture coefficients of the remaining leaf nodes with 1 iteration of ML estimation. Retraining was performed with all training utterances (no use of separate categories) since the aim was to elicit the interdependency between speaking rate and tree structure. Although the overall (line 'all') system performance does not change very much it can be observed that there is indeed a shift in recognition performance between the different rate categories. The improvements especially for the slow categories reach up to 11.8% relatively. However, for very fast speech the baseline system having more leaf nodes could not be outperformed.

5. DISCUSSION

In our paper we have shown several approaches for the generation of speaking rate dependent models. We proposed an approach based on a speaking rate feature included in the decision trees and a pruning technique suited for creating rate category dependent models. Experiments showed improvements, especially for extreme rate categories. Nevertheless for average speaking rate only slight improvements were achievable. Basically our experiments showed that fast speech tends to need more distinct tree nodes for modeling than slow speech.

6. ACKNOWLEDGEMENTS

This work was partially funded by the "Deutsche Forschungsgemeinschaft" (DFG).

7. REFERENCES

[1] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, M.A. Picheny, "Decision Trees for Phonological Rules in Continuous Speech", Proc. ICASSP 91, pp. 185-188.

[2] D.B. Paul, "Extensions to Phone-State Decision-Tree Clustering: Single Tree and Tagged Clustering", Proc. ICASSP 97, pp. 1487-1490.

[3] M. Finke, I. Rogina, "Wide Context Acoustic Modeling in Read vs. Spontaneous Speech", Proc. ICASSP 97, pp. 1743-1746.

[4] R. Kuhn, A. Lazarides, Y. Normandin, J. Brousseau, "Improved Decision Trees for Phonetic Modeling", Proc. ICASSP 95, pp. 552-555.

[5] A. Lazarides, Y. Normandin, R. Kuhn, "Improving Decision Trees for Acoustic Modeling", Proc. ICSLP 96, paper no. 158.

[6] I. Rogina, "Automatic Architecture Design by Likelihood-Based Context Clustering with Crossvalidation", Proc. Eurospeech 97, pp. 1223-1226.

[7] J.P. Verhasselt, J.-P. Martens, "A Fast and Reliable Rate of Speech Detector", Proc. ICSLP 96, pp. 2258-2261.

[8] N. Morgan, E. Fosler, N. Mirghafori, "Speech Recognition using On-Line Estimation of Speaking Rate", Proc. Eurospeech 97, pp. 2079-2082.

[9] N. Morgan, E. Fosler, "Combining Multiple Estimators of Speaking Rate", Proc. ICASSP 98, pp. 729-732.

[10] R. Faltlhauser, T. Pfau, G. Ruske, "On-line Speaking Rate Estimation Using Gaussian Mixture Models", Proc. ICASSP 2000, pp. 1355-1358.

[11] C. Fuegen, I. Rogina, "Integrating Dynamic Speech Modalities into Context Decision Trees", Proc. ICASSP 2000, pp. 1277-1280.

[12] W. Reichl, W. Chou, "A Unified of Incorporating General Features in Decision Tree Based Acoustic Modeling", Proc. ICASSP 99, paper no. 2377.

[13] T. Pfau, G. Ruske, "Creating Hidden Markov Models for Fast Speech", Proc. ICSLP 98, pp. 205-208.