Triphone Tying Techniques combining A–Priori Rules and Data Driven Methods

Ute Ziegenhain *Josef G. Bauer*

Siemens AG, Corporate Department Technology
81730 Munich, Germany
{Ute.Ziegenhain,Josef.Bauer}@mchp.siemens.de

Abstract

Tying of Hidden Markov Model states is an important issue for the use of triphones as modeling units in automatic speech recognition systems. This paper studies the application of a–priori rules for tying in combination with data driven methods. The baseline method features a combination of a–priori rules that reduce the theoretical number of units by an oder of magnitude and a simple back–off tying. Back–off tying is based on the frequency of units appearing in the training material. The use of the a–priori rules has practical advantages especially for the implementation of continuous phoneme recognition. This method is compared to the widely used decision tree based clustering that makes no use of a–priori rules. A third method is proposed that combines a–priori rules with decision tree based clustering. Experiments on telephone data show that the combined method outperforms both other methods preserving the advantages of applying a–priori rules.

1. Introduction

Most speech recognition systems use phonemes as their basic modeling units. When context independent phonemes are modeled the resulting number of parameters that have to be estimated is quite low. For European languages the typical number of phonemes is about 50. This results in a total of about 150 Hidden Markov Model states. This allows reliable estimations of parameters with limited training material.

To increase the acoustic resolution the use of context dependent phoneme modeling is an appropriate mean. The most popular kind of phonetic modeling is perhaps the use of triphones ([1], [2]). In this case the phoneme models depend on the preceeding as well as the succeeding phoneme. It is clear that in this case the theoretical number of models is several magnitudes higher than for context independent models. For the example of 50 phonemes the theoretical number of HMM states is about 3×50^3 when using three states per phoneme. In order to allow reliable estimation of model parameters some kind of parameter tying is crucial for triphone modeling. The goal of tying is the reduction of the number of HMM states by at least one order of magnitude.

In section 2 the baseline tying strategy with a–priori rules and simple back–off tying is described. The concept of decision tree based tying is explained in section 3. In section 4 a new method combining a–priori rules and decision trees is introduced. Experiments on telephone speech data are presented in section 5. Finally in section 6 some conlusions are drawn.

2. A–priori rules and simple back–off tying

In the described system a phoneme model consists of three Hidden Markov Model states. For a triphone state we use the following notation: $P_{current}(P_{left};P_{right}).idx$. In this case $P_{current}$ ist the current phoneme. P_{left} and P_{right} denote the preceeding and the succeeding phoneme. idx is in the range from 0 to 2 and denotes the state index. Our baseline tying method consists of a–priori tying rules and a simple back–off tying scheme that copes with unseen or rarely seen states.

2.1. A–Priori rules

The applied a–priori tying rules are as following:

1. All states with index 1 (central state) for a specific phoneme $P_{current}$ are tied together regardless of P_{left} and P_{right}. In other words: the central state of a phoneme is not context dependent.

2. All states with index 0 (left state) for a specific phoneme $P_{current}$ and specific phoneme P_{left} are tied together regardless of P_{right}. In other words: the left state of a phoneme depends only on the preceeding phoneme.

3. All states with index 2 (right state) for a specific phoneme $P_{current}$ and specific phoneme P_{right} are tied together regardless of P_{left}. In other words: the right state of a phoneme depends only on the succeeding phoneme.

The application of the given tying rules reduces the theoretical number of states dramatically. In the case of

N phonemes the theoretical number of states is $N \times N + N + N \times N = 2 \cdot N^2 + N$. For the example from above ($N = 50$) we end up with a theoretical number of units (HMM states) of 5050. For untied triphones this number would be 125000. It is clear that with the use of the a–priori rules the tying problem is partly solved.

Besides the partial solution of the tying problem there is another advantage of the applied a–priori tying rules. As only parts of the phoneme model depend on their context the rules can help to reduce the search space for decoding these models. Especially for continuous phoneme recognition the size of the required search space can be reduced considerably. Thinking of a static search space all middle states (index 1) $P_{current}(P_{left};P_{right}).1$. for a phoneme $P_{current}$ can be handled as one state in the search space. This would not be possible for general triphones as the context dependency of all states forbids path recombinations. For untied triphones a static search space must consist of $3 \times N^3$ states. If on the other hand the a–priori tying rules are applied the maximum size of the search space for continuous phoneme recognition is only $2 \cdot N^2 + N$ (states). In a real world system continuous phoneme recognition can be applied to realize a *say–in* application: the result of the phoneme recognition can be treated as a new entry in the phonetic lexicon for a new word entered with the users voice. Especially for embedded systems the reduced complexity for the continuous phoneme recognition can be necessary for the realization of an application.

2.2. Simple back-off tying

The second component of our baseline tying strategy is based on the frequency of the units in the training material. The following rules are applied:

- All left states $P_{current}(P_{left};P_{right}).0$ with a frequency below a certain threshold T are tied together regardless of P_{left} and P_{right}.

- All right states $P_{current}(P_{left};P_{right}).2$ with a frequency below a certain threshold T are tied together regardless of P_{left} and P_{right}.

A typical value for T is 100. Actually the described procedure results in some kind of *garbage* models. States with a low frequency are tied together with no respect to their acoustical similarity. But in practice this is far less harmful than it might appear because the frequency of this states in the test material also tends to be very low. A possibility for a more robust estimation of the garbage model would be the use of monophones that could be estimated from much more training material. This alternative was not investigated in the described work.

3. Decision tree based tying

Decision trees are a well known and often applied solution for tying problems resulting from context dependent phoneme models. The described system uses $3 \times N$ (3 states per phoneme, N phonemes) separate decision trees. To compute a decision tree we start with a pool consisting of all states $P_{current}(P_{left};P_{right}).idx$ for one value of $P_{current}$ and one value of idx. This pool corresponds to the root of the tree. This pool is then split into groups by adding new quesions to the decision tree. Firstly the pool at the root must be split. Further on more and more questions are introduced at some leaf of the tree so that the pools are split further on. The quesions asking for some phonetical properties of P_{left} and P_{right} split the pools in two pools of states for that the answer to the question is yes or no respectively. An example for such a question is 'Is P_{left} a vowel?'. The selection of the applied question is based on the log–likelihood gain achieved through splitting the pool at the node into two pools. The log–likelihood gain describes the improvement in the ability of the expanded tying specification to describe the training data. Another criterion for selection of the questions is the requirement that the total frequency (in the training material) of all states in a pool must be above a certain absolute threshold. Splitting is stopped when no further question can be found that would increase the log–likelihood by a certain threshold and fulfill the frequency criterion.

Once the tree is fully computed the tying specification for a triphone state is found by descending the tree while asking the questions at the nodes. The total number of all leaves (pools) in all trees then specifies the number of states with different probability density functions. To compute a set of trees with a certain total number of leaves we would just vary the log–likelihood and the frequency thresholds until an appropriate set of trees is found.

The clear advantage of the decision tree based tying is it's ability to take into account the acoustical similarities of states. Furthermore the concept of the phonetical questions has major advantages for unseen states. Although a certain state might not have been considered when computing the tree the phonetical questions apply to the unseen state.

4. Combining decision trees and a–priori rules

A new method is proposed now that combines the advantages of the a–priori tying rules and the concept of decision tree based tying using phonetical questions. We have developed a straightforward strategy to integrate the a–priori rules in the decision trees. For the computation of the decision trees the following rules must be applied:

- If idx is 0 (a left state in a phoneme) only questions

about P_{left} are allowed

- If *idx* is 1 (a central state in a phoneme) no questions at all are allowed
- If *idx* is 2 (a right state in a phoneme) only questions about P_{right} are allowed

These three rules make it very simple to create decision trees that incorporate the a–priori rules described in section 2.

The decision tree based tying considering a–priori rules preserves the advantages described in section 2. A further advantage that of course only applies for our systems is the possibility to reuse implementations based on the described baseline strategy.

5. Experiments

5.1. Baseline system

All experiments are performed on telephone quality speech data at a sample rate of 8 kHz. The feature extraction is based on cepstral features and an energy parameter along with first and second order derivatives. Online spectral mean and energy mean removal are applied. Supervectors consisting of several cepstral feature vectors are transformed using a linear transformation matrix based on Linear Discriminant Analysis (LDA). The final feature vector is formed from the 24 most discriminant components of the transformed vector. Details on the feature extraction can be found in [3].

The recognition systems used for the described experiments is based on Continuous Density Hidden Markov Models. State specific probability density functions are realized via mixtures of Gaussian densities. For recognition purposes unified variance modeling with only one global variance parameter is applied. For computation of the decision trees diagonal variance matrices are estimated. As already noted each phoneme model consists of three states. The states are in Bakis–topology with only self loop, one state transition and skips being allowed. For parameter estimation Viterbi based Maximum Likelihood training is performed. Similar systems are described in [3] and [4].

For the phoneme based models the phoneme inventory consists of 48 different phonemes. For computation of the decision trees the applied questions are formed by 132 questions about phonetical features and $48 \times 2 = 96$ questions about the particular phonemes itself.

5.2. Experiments on natural numbers

In a first series of experiments training and tests are carried out on parts of the German SpeechDat databases ([5]) containing natural numbers in telephone quality. At a vocabulary size of about 115 it is clear that in this case no problems with rarely seen states occur. Nevertheless tying is an important issue for this application since real world systems like embedded systems can have a limitation on the number of different states as low as 256. For this task we compared context independent phoneme models, untied triphones, triphones with a–priori tying rules and triphones with decision tree based tying. A total number of 4000 Gaussian densities was chosen for all models.

From the results in table 1 it can be seen that (untied) triphones clearly outperform context independent monophones. The total number of states for the untied triphones is quite high at 342. Application of the a–priori tying rules reduces the number of states to 175 while the error rate increases only very slightly. When using decision trees to reduce the number of states to the same amount we observe a bigger increase of the error rate.

context	tying	#states	WER
monophone	—	84	12.6
triphone	—	342	8.3
triphone	APR	175	8.5
triphone	DT	175	9.0

Table 1: Word error rates (WER) for different phonetical modeling and tying, APR: a–priori rules, DT: decision trees

This experiment shows that the assumptions on which the a–priori tying rules are based are at least partially fulfilled here. Data driven tying by decision trees could have produced exactly the same tying rules but yielded to a significantly worse partitioning of the HMM states in terms of word error rate. This is possible as the log–likelihood criterion for construction of the decision trees is not perfectly correlated with the word error rate.

5.3. Experiments with training on phonetically rich material

For a second series of experiments only phonetically rich speech material from the German SpeechDat II database served as training material for a non task specific set of Hidden Markov models. From the theoretical number of 110592 possible triphones only about 12600 occur in the training corpus. For this kind of applications tying must be able to deal with rarely seen or unseen states.

In a first step the a–priori tying rules and simple backoff tying (threshold 150 for frequency of states) were applied. This leads to a number of 1780 different triphone states. Two other possibilities for triphone tying based on decision trees were investigated. First standard decision trees were computed. Secondly decision trees incorporating the a–priori tying rules according to section 4 were computed. For both cases the total number of states

was chosen to 1780 for comparison purposes. All models consist of about 20000 Gaussian densities.

For evaluation of the recognition performance three different tasks were considered: isolated command words, continuous digits and continuous spelling. The isolated word tasks with 115 different words contains a number of words with rarely seen triphones. Therefore this task is especially important for evaluation of different tying strategies.

Table 2 shows the word error rate for the different tying strategies and recognition tasks. It can be seen that

tying	WER-COM	WER-CD	WER-SP
APR+SBO	2.4	7.4	39.9
DT	2.1	6.2	37.7
DT+APR	1.5	6.5	36.1

Table 2: Word error rates (WER) for different taks and tying strategies, APR: a–priori rules, SBO: simple back-off, DT: decision trees, ER-COM: error rate on isolated command words, ER-CD: error rate on continuous digits, ER-SP: error rate on cont. spelling

for this tasks standard decision trees (DT) clearly outperform the use of a–priori tying rules and simple back-off (APR+SBO). But it turns out that the combination of decision trees and the a–priori rules (DT+APR) results in the overall best recognition results. Only for the continuous digit task DT+APR performs slightly worse than DT.

The assumptions on that the a–priori rules are based proof to be valid here, too. The unconstrained tying according to the a–priori rules does not seem to cause some loss on modeling accuracy. Comparing all three results it appears that it is the simple back-off tying for rarely seen states that impairs the recognition performance for the APR+SBO system. When comparing the DT versus the DT+APR tying strategy in detail it turns out that the major difference is the structure of the trees for the central states (*idx* 1). While for the DT method these trees contain usually several ten leaves the trees considering the a–priori rules have only one leaf. This implies a higher resolution for the left (*idx* 0) and the right (*idx* 2) states of the triphones. It seems that this enhanced resolution is of greater importance than the resolution for the middle states.

6. Conclusions

Three methods for tying of triphone HMM states were investigated. For two methods a–priori tying rules were applied that cause tying of all middle triphone states for a specific phoneme and consider only the left / right context for left / right states of triphones. A practical advantage of this rules is a reduced complexity for decoding context dependencies across units. This is especially useful for continuous phoneme recognition with applications for speaker dependent recognition.

In the context of natural number recognition tying based on these rules was compared to tying based on the commonly used decistion trees. It was found that in this case the a–priori rules based tying outperformed the decistion trees. This leads to the assumption that the a–priori rules are suitable for triphone tying in automatic speech recognition and can outperform other well established methods.

A second set of experiments considers HMM training on phonetically rich speech data only. In this context the rarely seen and unseen triphones are of great importance. Decision trees were compared to a–priori rules with a simple back–off strategy mainly based on the frequency of states as seen in the training material. Recognition results on different tasks showed that the decision tree based tying outperforms the other method. A newly proposed method combining decision trees and a–priori rules was also compared to the other two methods. This new method was found to outperform both other methods.

It can be concluded that the a–priori tying rules are very suitable for tying of triphone HMM states but should be combined with decision trees to improve handling of rarely seen states which is not treated efficiently by the simple back–off tying. A method was proposed that allows integration of the a–priori rules in the computation of the decision trees very easily.

7. References

[1] S.J. Young and P.C. Woodland, "The use of state tying in continuous speech recognition," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 1993, pp. 2203–2206.

[2] K. Beulen, E. Bransch, and H. Ney, "State tying for context dependent phoneme models," in *ICASSP*, 1997, pp. 1179–1182.

[3] Josef G. Bauer, "Enhanced control and estimation of parameters for a telephone based isolated digit recognizer," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, pp. 1531–1534.

[4] Josef G. Bauer and Jochen Junkawitsch, "Accurate recognition of city names with spelling as a fall back strategy," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 1999, vol. 1, pp. 263–266.

[5] "Elra web site," http://www.icp.grenet.fr/ELRA.