# Some Investigations of the Decision Trees from the ASR, TTS and Applied Phonetics Point of View

**Stefan Grocholewski**
Poznan University of Technology, Poznan, Poland
E-mail: stefan.grocholewski@cs.put.poznan.pl

## ABSTRACT

In our HMM's based ASR system we have used the models of Polish triphones. To deal with triphones for which there are no examples in the training data we introduced decision tree-based clustering. In the paper we tried to analyze the clusters with the identical triphones and the questions set forming the decision trees obtained in our experiments with the ASR system. The results of our investigations of the equivalent contexts leading to the acoustically similar triphones (they share common clusters for all states) should be useful from the point of view of speech recognition, text to speech systems and applied phonetics.

## 1. INTRODUCTION

The decision tree-based states clustering mechanism offers a solution to the unseen triphone problem. A phonetic decision tree is a tree in which a binary phonetic question is attached to each node. Initially all states are placed in a single cluster at the root of the tree. The question at each node is chosen in order to maximize the likelihood of the training data. Splitting any node into two will increase the log likelihood. The increase obtained when each possible question is used can be calculated and the "best" question selected which gives the biggest improvement [3]. The process of node splitting is stopped if the log likelihood gain becomes lesser than the predefined threshold. The states in each cluster are then tied. After tying procedure we can obtain many identical triphones, with identical all appropriate states. In this paper we will study which contexts are equivalent, i.e. result in the same states. It will help us to compare our theoretic phonetic knowledge with the real data. We will also investigate the sequences of questions in the decision trees since their succesion is connected with their weights. In the section 2 we will present the database for Polish used in our experiments and in the section 3 the details about the acoustic modeling will be presented.

## 2. DATABASE FOR POLISH

The investigation has been made on the base of the first database for Polish CORPORA [2]. It was created to comprise the relatively large number of Polish diphones. The final number of diphones was equal to 179.753 and the number of different diphones we have found in CORPORA was equal to 1271. Note that 37 Polish phonemes implies about 1440 theoretically possible diphones. The database is distributed with the ASCII files in HTK [3] format comprising the phonemic transcriptions along with the timing data.

## 3. ACOUSTIC MODELING OF POLISH

We have chosen the context dependent phonemes (triphones) as the phonetic units and Continuous Density Hidden Markov Models (CDHMM) as the models. For training models of 37 Polish phonemes we have used 5110 utterances of CORPORA (14 male speakers x 365 utterances). We found 2953 different triphones which give 8859 states (3 states per model). The biggest problem in building context dependent HMM's is data insufficiency. Considering the size of our database we can expect that not all triphones will be reliably estimated. The solution consists in tying the states within triphone sets in order to share data and thus be able to make robust parameter estimates. The choice of which states to tie requires a little subtlety since the performance of the recognizer depends crucially on how accurately the state output distributions capture the statistics of the speech data. In our case we have chosen the decision tree based approach [3] to tie the similar states. This approach helps to solve the problem of the unseen triphones, i.e. it allows triphone models which have no training data to be synthesized. The phonetic decision trees are built using a top-down sequential optimization process. Initially all states are placed in a single cluster at the root of the tree. The question is then found which gives the best split of the root node. This process is repeated until the increase in log likelihood falls below the specified threshold. As a final stage, the decrease in log likelihood is calculated for merging terminal nodes with different parents. Any pair of nodes for which this decrease is less than the threshold used to stop splitting are then merged.

In the section 4 we will investigate the sequence of questions during the decision tree construction.

After tying procedure we can expect that different triphones from the same monophone can share the same corresponding states. In the section 5 we will investigate the identical triphones. It will be interesting to know

 ISBN 1-876346-48-5 

which contexts are equivalent (only for our database of course) and if this knowledge corresponds to the acoustic phonetic theory.

## 4. ANALYSIS OF THE QUESTIONS IN THE DECISION TREES

In three states models, in very coarse approximation, we assume that the left state is left context dependent, the middle one is context independent and the right one is right context dependent. To verify this opinion in the Table 1 we present the results of counting the questions about the appropriate context. We can see, for example, for the vowel *i* that the left state always (for our database) depends only on the left context (100%) whereas the right state of the model depends not only on the right context (in 70,8%) but also on the left one. Nevertheless, studying the Table 1 we can state that in most cases the boundary states depend also on the opposite contexts.

| | phoneme | questions about context (in %) | | | |
|---|---|---|---|---|---|
| | | state1 | state 2 | | state 3 |
| | | left | left | right | right |
| 1 | *i* | 100 | 34,5 | 65,5 | 70,8 |
| 2 | *ɨ* | 100 | 65 | 35 | 73,7 |
| 3 | *e* | 85,2 | 50,9 | 49,1 | 75 |
| 4 | *a* | 84,6 | 46,9 | 53,1 | 90 |
| 5 | *o* | 88 | 51,9 | 48,1 | 92,9 |
| 6 | *u* | 86,7 | 53,3 | 46,7 | 80 |
| 7 | *õ* | 90 | 77,8 | 22,2 | 36,4 |
| 8 | *ẽ* | 50 | 28,6 | 71,4 | 75 |
| 9 | *w* | 60 | 71,4 | 28,6 | 58,3 |
| 10 | *j* | 88,2 | 75 | 25 | 45,4 |
| 11 | *l* | 60 | 52,6 | 47,4 | 50 |
| 12 | *t* | 87,5 | 45,5 | 54,5 | 50 |
| 13 | *p* | 64,3 | 80 | 20 | 70 |
| 14 | *k* | 70 | 46,2 | 53,8 | 54,5 |
| 15 | *d* | 100 | 80 | 20 | 88,9 |
| 16 | *b* | 53,8 | 14,3 | 85,7 | 100 |
| 17 | *g* | 63,6 | 50 | 50 | 100 |
| 18 | *ts* | 87,5 | 42,9 | 57,1 | 50 |
| 19 | *ʧ* | 100 | 50 | 50 | 75 |
| 20 | *tɕ* | 88,9 | 20 | 80 | 50 |
| 21 | *dz* | 0 | 0 | | 0 |
| 22 | *dʑ* | 100 | 0 | 100 | 100 |
| 23 | *dʒ* | 100 | 0 | 100 | 100 |
| 24 | *m* | 55,6 | 52,9 | 47,1 | 66,7 |
| 25 | *n* | 87,5 | 40 | 60 | 72,7 |
| 26 | *ɲ* | 100 | 71,4 | 28,6 | 66,7 |
| 27 | *ɳ* | 100 | 100 | 0 | 100 |
| 28 | *s* | 90 | 42,1 | 57,9 | 100 |
| 29 | *ʃ* | 71,4 | 55,6 | 44,4 | 71,4 |
| 30 | *ɕ* | 100 | 66,7 | 33,3 | 85,7 |
| 31 | *f* | 80 | 83,3 | 16,7 | 75 |
| 32 | *x* | 83,3 | 50 | 50 | 75 |
| 33 | *r* | 72,7 | 50 | 50 | 58,3 |
| 34 | *w* | 50 | 33,3 | 66,7 | 85,7 |
| 35 | *z* | 87,5 | 40 | 60 | 87,5 |
| 36 | *ʑ* | 100 | 33,3 | 66,7 | 83,3 |
| 37 | *ʒ* | 100 | 100 | 0 | 100 |
| | | **83,0** | **51,5** | **48,5** | **75,4** |

**Table 1:** Results of counting the questions for all states about the left or right context

We have found some very interesting exceptions. Splitting the left state of the consonant *v* on the first step, before asking about the left context, we should ask if the right context is *i*. We have found the same situation splitting the right state of the fricative *ʃ*. The main question was about the left context (is it *u*?). A similar case was for fricative *f*. Splitting the right state we should ask if the left context is *a*. To complete the exceptions we should mention the right state of the plosive *k* (depends mainly on the *e* on the left side) and *w* - where the left state depends mainly on the *a* on the right side.

When interpreting the above results we have been conscious that there was a dependence between the implementation of the particular question for the particular phone and the number of data observations for that phone. If the database used for the design of the phone set has phone frequencies which are very different from the actual recognition system environment, the resulting decision tree and the designed phone set may be different.

| | Number of occurrences | Number of different triphones | | Number of occurrences | Number of different triphones |
|---|---|---|---|---|---|
| *a* | 5857 | 222 | *b* | 966 | 65 |
| *e* | 4043 | 220 | *f* | 942 | 50 |
| *o* | 3368 | 192 | *ʃ* | 862 | 61 |
| *n* | 2561 | 114 | *g* | 789 | 56 |
| *j* | 2205 | 98 | *ts* | 779 | 57 |
| *ɨ* | 2193 | 127 | *z* | 772 | 67 |
| *r* | 2070 | 103 | *ʒ* | 759 | 67 |
| *m* | 1914 | 108 | *x* | 740 | 53 |
| *t* | 1828 | 80 | *õ* | 715 | 59 |
| *l* | 1772 | 86 | *ɕ* | 661 | 66 |
| *ɨ* | 1722 | 105 | *tɕ* | 588 | 52 |
| *u* | 1689 | 97 | *ʧ* | 508 | 43 |
| *k* | 1514 | 79 | *ʑ* | 450 | 43 |
| *w* | 1447 | 83 | *dʑ* | 383 | 37 |
| *s* | 1319 | 86 | *ẽ* | 315 | 37 |
| *w* | 1235 | 73 | *dʒ* | 247 | 23 |
| *d* | 1167 | 85 | *dz* | 160 | 18 |
| *p* | 986 | 59 | *ɳ* | 108 | 16 |
| *ɲ* | 967 | 64 | **Σ** | **52418** | **2953** |

**Table 2:** Statistics of the dataset used in the experiments

In the table 2 we present the statistics of our dataset; the number of occurrences of the particular phonemes along with the number of different triphones for that phoneme.

In the table 3 we present the most important questions about the left and right context for all phonemes.

| | | left state | right state |
|---|---|---|---|
| 1 | *i* | *ɲ,m,l,dʑ,v* | *w,m,n,k,o,r,z* |
| 2 | *ɨ* | *m,w,n,r, ʒ,b* | *w,,n,m,j,s,k,b* |
| 3 | *e* | *j, ɲ,l, dʑ, tɕ, ʒ* | *j,n,m, ɲ,w,o,r* |
| 4 | *a* | *j,n,w,m,l,r,d* | *n,j,r,w,l,m* |
| 5 | *o* | *r,n,j,e,d,w* | *ɲ,n,l,j,m,k,s,r* |
| 6 | *u* | *j,m,l,e,r,g* | *j, ʃ,l,r, ʑ,* ***ts*** |
| 7 | *õ* | *j,n,d,s,* ***ʒ*** | *e,u, ʒ,i,****s,d,j*** |
| 8 | *ẽ* | *ɕ,****s*** | *s, ʃ,* ***ɕ,****z* |
| 9 | *w* | ***a****,a,e, ɨ,i* | *a,o, ɨ,u,g,z,****u*** |
| 10 | *j* | *a,e,u,m,o,r* | *a,o,u,o, õ,****ɲ*** |
| 11 | *l* | *a,o,e,u,i,****i*** | *e,a,i,u,j,o,****a,o*** |
| 12 | *t* | *a,o,s,r,e,n* | *a,o,e, ɨ,****s,n,r,e*** |
| 13 | *p* | *a,o,u* | *a,o,i,j* |
| 14 | *k* | *i,o,e,a, ɨ* | *a,****e****,o,****i****, ****ɨ**** |
| 15 | *d* | *o,a,i,e* | *a,o,e, õ ,j* |
| 16 | *b* | *o,a,u,r,****e,e*** | *a,i,e,o,e,r* |
| 17 | *g* | *a,o,i,w,****e,u*** | *u,o,a,e,i* |
| 18 | *ts* | *a,i,o,u,****a*** | *j,e, ɨ,a* |
| 19 | *tʃ* | *e, ɨ,i,a* | *ɨ,e,a,****i*** |
| 20 | *tɕ* | *ɕ,a,o,****a*** | *i,e,a,****i****, ****ɲ****,****a*** |
| 21 | *dz* | - | - |
| 22 | *dʑ* | *o,a,i* | *i,e* |
| 23 | *dʒ* | *e* | *o,e* |
| 24 | *m* | *e,a,o,i, ɨ,****a****,****i****,****u*** | *a,i,j,u, ɨ* |
| 25 | *n* | *a,o,e,i, ɨ,n* | *a,t,o, õ,e,n* |
| 26 | *ɲ* | *o,e,a* | *e,i, tɕ,j,****e,o*** |
| 27 | *ɳ* | *o* | *g* |
| 28 | *s* | *a,o,e, ɨ* | *w,t,k,e* |
| 29 | *ʃ* | *u,a,e,i, ɨ,****ɨ*** | ***u****, ɨ,a,e,k,****a*** |
| 30 | *ɕ* | *i,a,o,e,m* | *tɕ,i,e, ẽ* |
| 31 | *f* | *a,e,o,u,****e*** | ***a****,a,j,i,p,****e*** |
| 32 | *x* | *e,a,iu,o,****u*** | *a,o,u,w,****e,u*** |
| 33 | *r* | *a,u,e,o,i,****u*** | *t,o,a,j,u* |
| 34 | *v* | ***i****,e,o,a,****j****,d,****u,b*** | *a,i,o,j,u,****o*** |
| 35 | *z* | *i,a, ɨ,o* | *a,j,e, ɨ,o* |
| 36 | *ʑ* | *u,a,o* | *i,e,u,****o*** |
| 37 | *ʒ* | *a,o,d* | *ɨ,e,a,o* |

**Table 3:** The most important questions about the left and right context for all phonemes

The very general conclusion is that for vowels the most important questions were about nasals, liquids and glides whereas for consonants the most important questions were about vowels. Letters in bold in the Table 3 mean that an appropriate state depends on the opposite context - see the beginning of the section.

## 5. EQUIVALENT CONTEXTS

After tying procedure for almost all phonemes we have obtained some clusters of triphones. In every cluster all triphones have identical appropriate states. In the table 4 the total number of different triphones, the number of clusters, the number of triphones outside clusters and the number of triphones per cluster are presented.

| | different tripnones | number of clusters | outside clusters | triphones per cluster |
|---|---|---|---|---|
| *a* | 222 | 10 | **87%** | 2,8 |
| *õ* | 59 | 13 | 12% | 4 |
| *b* | 65 | 11 | 28% | 4,3 |
| *ts* | 57 | 8 | 28% | 5,1 |
| *tʃ* | 43 | 0 | **100%** | |
| *tɕ* | 52 | 11 | 10% | 4,3 |
| *d* | 85 | 11 | 40% | 4,6 |
| *e* | 220 | 19 | **77%** | 2,8 |
| *ẽ* | 37 | 6 | 11% | 5,5 |
| *f* | 50 | 10 | 8% | 4,6 |
| *g* | 56 | 8 | 37% | 4,4 |
| *x* | 53 | 10 | 11% | 4,7 |
| *i* | 127 | 20 | 42% | 3,7 |
| *j* | 98 | 10 | 37% | 6,2 |
| *k* | 79 | 11 | 39% | 4,4 |
| *l* | 86 | 9 | 35% | 6,2 |
| *w* | 83 | 13 | 13% | 5,5 |
| *m* | 108 | 13 | 21% | 6,5 |
| *n* | 114 | 13 | 27% | 6,4 |
| *ɲ* | 64 | 6 | 16% | 9 |
| *o* | 192 | 19 | **78%** | 2,3 |
| *p* | 59 | 16 | 17% | 3,1 |
| *r* | 103 | 13 | 18% | 6,5 |
| *ʒ* | 67 | 10 | 19% | 5,4 |
| *s* | 86 | 12 | 38% | 4,4 |
| *ʃ* | 61 | 10 | 20% | 4,9 |
| *ɕ* | 66 | 13 | 14% | 4,4 |
| *t* | 80 | 13 | 36% | 3,9 |
| *u* | 97 | 19 | 43% | 2,9 |
| *w* | 73 | 10 | 14% | 6,3 |
| *ɨ* | 105 | 19 | 45% | 3,1 |
| *z* | 67 | 14 | 18% | 3,9 |
| *ʑ* | 43 | 6 | 12% | 6,3 |
| *dz* | 18 | 0 | **100%** | |
| *dʒ* | 23 | 0 | **100%** | |
| *dʑ* | 37 | 0 | **100%** | |

**Table 4:** The total number of different triphones, the number of clusters, the number of triphones outside clusters and the number of triphones per cluster


ISBN 1-876346-48-5 

For four consonants: *tʃ, dz, dʒ, dʑ* we have observed no clusters. It means that they are very context dependent. Almost the same can be said about three vowels: *a*, *e*, *o* (about 80 % outside clusters). On the opposite side we have diphthongs: *õ*, *ẽ* and fricatives *f*, *x*.

It is interesting to study the lists of the identical triphones and triphones placed in different clusters. For example, the set {*i+b*, *x-i+b*, *ts-i+dʑ*, *i+dʒ*, *i+dʑ*, *i+e*, *i+f*, *i+p*, *i+v*, *j-i*, *t-i+dʒ*, *r-i*, *n-i+e*, *a-i+dʑ*, *i-i+p*} forms one cluster, whereas triphones *i+m* and *i+n* form different clusters. The notation h-i+b describes the phoneme i with the left context *x* and right context *b* [3]. We should be careful since our tying procedure consists in tying the states in each leaf node in the decision tree. It means that the triphones in one cluster are the same but at the same time the triphones in different clusters could be the same. For example, *d-a+m* and *d-a+n*, *n-a+m* and *n-a+n* are not tied because the questions about *m* and *n* are the subsequent questions building the decision tree and they must be in different clusters. We noticed the same situation for all vowels. Thus, one can suggest asking not about the particular phonemes but about the classes as in [1]. Nevertheless, in such a solution we are not able to investigate the succession of questions. In the question about the right state of *e* the nasals are on the: second (n), third (m) and fourth (ɲ) position whereas in the question about the right state of *a* the nasals *m* and *n* are separated by five other questions. Considering one cluster we can draw conclusions about the equivalent contexts. For example in the cluster {*n-a+a*, *n-a+e*, *n-a+i*, *n-a+o*, *n-a*} the right vowel contexts are not distinguishable.

## 6. DISCUSION OF RESULTS

Analyzing the questions in the decision trees and the clusters of triphones obtained after tying the states in leaves we can enhance our knowledge of acoustic phonetics from the statistical point of view. With this technique we can discover new dependencies, especially connected with the significance of the context. This knowledge can be useful from the ASR, TTS and applied phonetics point of view. The obtained results depend on the dataset. To confirm our observations we analyzed the appropriate decision trees for another dataset with different vocabulary. The results were very similar. It is impossible to present all of them. The results are in the form of the set of questions and the contents of the leaves in the decision trees. The goal of our paper was to present rather the used “technology” than the detailed results.

## 7. CONCLUSION

In the paper the investigations of the decisions trees used in our HMM based ASR system have been presented. Two kinds of data: the succession of questions and the contents of leaves have been studied. In our opinion the results of the investigation should enhance our phonetic knowledge from the ASR, TTS and applied phonetics point of view.

## ACKNOWLEDGMENT

This work was supported by KBN, grant no. 7T 11C 009 21

## REFERENCES

[1] V.Chuchupal, K.Makovkin, A.Chichagov, Accurate acoustic modelling for Russian, *Proc. SPECOM’2000*, pp.71-74, 2000.

[2] S.Grocholewski, First Database for Spoken Polish, *Proc. Int. Conf. on Language Resource & Evaluation*, pp.1059-1062, 1998.

[3] S.Young, at al., *HTK Book*, Cambridge University, 1997.