MUTUAL INFORMATION PHONE CLUSTERING FOR DECISION TREE INDUCTION

Ciprian Chelba

Microsoft Research
One Microsoft Way
Redmond, WA 98052

Rachel Morton

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

ABSTRACT

The paper presents an automatic method for devising the question sets used for the induction of classification and regression trees. The algorithm employed is the well-known mutual information based bottom-up clustering applied to phone bigram statistics. The sets of phones at the nodes in the resulting binary tree are used as question sets for clustering context-sensitive (tri-phone) HMM output distributions in a large vocabulary speech recognizer. The algorithm is shown to perform as well and sometimes significantly better than question sets devised by human experts for a Spanish and German system evaluated on several tasks, respectively. It eliminates the need for linguistic expertise and it provides a faster solution as well.

1. INTRODUCTION

State-of-the-art speech recognition technology uses phone level HMMs to model the speech feature vector produced when uttering a sequence of words. The conversion from words to phones is accomplished by pronunciation dictionary look-up. Using context dependent HMM models for each phone results in better acoustic models: in a tri-phone system, the phone to the left/right of the current phone is also taken into account and we thus use an HMM for each tri-phone. This results in an exponential increase of the number of parameters in the acoustic model with the length of the context, leading to data sparseness issues when estimating the model parameters. In order to balance the modeling accuracy brought by increased context length with the reliability of the parameter estimates, decision tree clustering of tri-phone states/output distributions has been widely used, [1], [2], [3]. The decision tree uses sets of phones to whom membership of the current/left/right phone is ascertained at various nodes in the tree.

The standard approach to deriving the question sets for a particular language is to use a human expert. For some languages the linguistic expertise might not exist at all, be scarce or expensive. We propose a simple automatic procedure for inferring the question sets that uses solely text (word strings) in the language of interest and a pronunciation dictionary. The approach eliminates the need for a human expert. In the same vein of thinking, however using a completely different approach, [4] tries to discover question sets for decision-tree-based letter-to-sound rules.

Our algorithm uses the phone strings obtained by mapping the words in the text transcription through the pronunciation dictionary to extract bigram phone co-occurence statistics. The well-known mutual information-based clustering algorithm [5] is then employed to derive a binary clustering tree for phones. Each node in the tree will contain a set of phones which are retained as the automatically inferred question sets.

The question sets have been evaluated against manually derived question sets used for building clustering trees for recognition systems in German and Spanish. For each language, we have built two systems, one using the manually derived question set and one using the automatic one. The clustering trees were grown such that the number of clustered states was similar. Several test sets were used to evaluate each system. The automatic question set performed about the same as the manual one, sometimes significantly better.

The approach offers an attractive alternative when deploying a speech recognizer in a new language, making the localization process easier. Other potential applications are for decision tree-based letter-to-sound rules used when generating pronunciations for unknown words in speech recognition/synthesis.

The paper is organized as follows: Section 2 gives an overview on the HMM state clustering procedure employed by the HTK training tools [6]. Section 3 describes the mutual information clustering algorithm used for question set induction and Section 4 reports the experimental results. We conclude with Section 5.

2. OVERVIEW OF HMM STATE CLUSTERING PROCEDURE

In order to effectively balance the model size with the reliability of estimates in a tied-state HMM system, state based

clustering is usually employed. The most advantageous method has been proven to be the top-down decision tree based clustering, [1], [2], [3]. It is computationally cheaper than bottom-up clustering and it also deals gracefully with contexts (triphones) that have not been encountered in the training data.

As explained in [3], the process of building a tied state HMM recognition system using HTK proceeds as follows:

1. train an initial system that uses context independent phones (monophones); each phone HMM has a fixed number of states, typically 3, and each state has a single Gaussian output probability density function (pdf)
2. "clone" each monophone encountered in the training data into a triphone; all triphones that have the same central phone end up with different HMMs that have the same initial parameter values; train the resulting triphone system
3. cluster corresponding states in the triphone HMMs that share the same central phone
4. increment the number of Gaussian components for each tied state output pdf and re-train the system until the performance on a development set peaks

The third step employs a phonetic decision tree for pooling triphone states into equivalence classes that will each use a separate output pdf. The decision tree asks whether the phone to the left/right of the central phone is in a certain set, e.g. "Is the phone to the left/right a vowel?" — these are the question sets we are seeking to derive in an automatic fashion.

For details on the decision tree induction the reader is refered to [3]. Overly simplifying the procedure, one could describe it as:

1. one tree is constructed for each state of all triphones that share the same central phone, e.g. state 2 of d-eh+ih" and state 2 of m-eh+g" will be using the same clustering tree, whereas states 2 and 3 of the same triphone will not; if $\mathcal{P}$ is the phone vocabulary and each phone HMM has S states, we end up building $\mid \mathcal{P} \mid \cdot S$ trees
2. each tree is built top-down such the likelihood of the training data is greedily increased at each split; a possible stopping criterion is to stop splitting a given node when the likelihood increase resulting from that split falls below a given threshold
3. the splits are determined by asking questions about the set membership of phones to the left/right of the central one

At test time, unseen contexts are handled naturally by the above clustering tree since it induces a disjoint partition on the set of all possible triphones.

The question sets used for determining the splits in the decision tree are usually devised by a linguist expert. Examples for English are: "Vowel", "Unrounded", "UnFortisLenis", "Fortis", etc. We propose an approach that derives them automatically.

3. AUTOMATIC INDUCTION OF QUESTION SETS

The question sets used for building the state clustering tree are built as follows:

1. obtain phone transcriptions using Viterbi aligments of context independent models (used for picking the best pronunciation for words that have multiple pronunciations)
2. gather phone co-occurence bigram statistics
3. run bottom-up mutual-information clustering algorithm [5] to derive a binary tree; each node in the tree except for the root node — including the leaves, containing exactly one phone — is retained as a question set

For the sake of completeness we reproduce here the basic idea behind the mutual-information clustering algorithm [5].

Let $\mathcal{P}$ be the phone vocabulary. We denote by $f(p_1, p_2)$ the bigram relative frequency estimate of phone p_2 following phone p_1. Assume that at a given stage during the bottom-up clustering process we have a set $\mathcal{C}$ of n mutually disjoint phone clusters

$$\mathcal{C} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n; \mathcal{S}_i \cap \mathcal{S}_j = \emptyset, \cup_{i=1}^{n} \mathcal{S}_i = \mathcal{P}\}$$

Let $f_{\mathcal{C}}(\mathcal{S}_i, \mathcal{S}_j) = \sum_{p_1 \in \mathcal{S}_i, p_2 \in \mathcal{S}_j} f(p_1, p_2)$ denote the cluster bigram relative frequency induced by $f(p_1, p_2)$ and a given set of clusters $\mathcal{C}$. Based on it one can calculate the mutual-information between adjiacent clusters as:

$$MI_{\mathcal{C}}(S_-, S_+) = \sum_{\substack{i,j=1\ldots n \\ i \neq j}} f_{\mathcal{C}}(\mathcal{S}_i, \mathcal{S}_j) \log \frac{f_{\mathcal{C}}(\mathcal{S}_i, \mathcal{S}_j)}{f_{\mathcal{C}}(\mathcal{S}_i) \cdot f_{\mathcal{C}}(\mathcal{S}_j)}$$

We seek to merge two clusters $\mathcal{S}_i$ and $\mathcal{S}_j$ resulting in a new set of clusters $\mathcal{C}'$ such that the mutual information loss is minimized[1]:

$$\max_{\mathcal{C}'} \triangle MI(S_-, S_+) = MI_{\mathcal{C}'}(S_-, S_+) - MI_{\mathcal{C}}(S_-, S_+)$$

Equivalently, one looks for the set of clusters $\mathcal{C}'$ that yields the maximum mutual information $MI_{\mathcal{C}'}(S_-, S_+)$.

The algorithm is initialized by placing each phone in a cluster of its own. The above cluster merging strategy is applied repeatedly until all phones are in one cluster. All the

[1] $\triangle MI(S_-, S_+)$ is a negative quantity because each merge is a partition of $\mathcal{C}$ that reduces the KL-divergence $D(f(\mathcal{S}_i, \mathcal{S}_j) \parallel f(\mathcal{S}_i) \cdot f(\mathcal{S}_j))$

nodes in the tree except for the root are retained as question sets.

The complexity of the algorithm is $\mid \mathcal{P} \mid^3$ but since the size of the phone vocabulary is typically less than 100 the algorithm is computationally tractable without resorting to the approximations described in [5] for large vocabularies.

3.1. Why would any of this work?

A legitimate question at this point is why would the mutual information based clustering algorithm produce reasonable questions to be used in the phonetic decision tree?

The above procedure is usually employed for building cluster-based language models and it will yield clusters that are optimal for predicting the identity of the current phone given the cluster of the left/right one, as explained in [5]. Assuming that the phone set is designed such that it carries as much acoustic information as possible, i.e. the phone set is such that various labels correspond to different acoustic realizations of a given phone, the above clusters will thus be relevant to the acoustic realization of the central phone.

4. EXPERIMENTAL RESULTS

The mutual information based questions were evaluated using an acoustic model set consisting of tied-state cross-word triphone HMMs. To evaluate whether the method extends to new languages the experiment was carried out for German and for Spanish.

4.1. Training Mutual Information Based Question Sets

Automatic questions were produced for German and Spanish and compared to hand-written questions for each language.

The procedure for training automatic questions was as follows. A twelve-mixture monophone HMM set which had never previously been clustered was used to perform a Viterbi alignment of the training transcriptions against a pronunciation dictionary. This was performed by the HTK tool HVite[6]. The German input lexicon contained 51 phones and 19,132 unique words. The Spanish input lexicon contained 28 phones and 11,501 unique words. At each point in alignment the best matching alternate pronunciation was selected. The alignment output was then converted to a pronunciation lexicon which was used as input to train the mutual information based questions.

The German training data consisted of 86781 utterances from 2045 different speakers from the SpeechDatM (SpM) corpus, and the SpeechDatII (SpII) corpus, available from ELRA[7]. The Spanish training data consisted of 71960 utterances from 1717 different speakers of the SpeechDatM Castillian Spanish corpus and the VAHA Polyphone Hispanic Spanish corpus, available from the LDC[8].

```
QS L_20 { l-*,m-*,ng-* }
QS L_22 { tS-*,pf-*,s-*,S-*,Z-* }
QS L_23 { OY-*,U-* }
QS L_24 { h-*,j-*,v-* }
QS L_25 { a:-*,e:-*,E:-*,E6*,2:-*,
o:-*,u:-*,y:-* }
QS L_26 { 6-*,x-* }
QS L_27 { aI-*,aU-* }
QS L_38 { a-*,a:-*,a6-*,aI-*,O-*,
aU-*,e:-*,E-*,E:-*,
E6-*,2:-*,i:6-*,o:-*,
oe-*,u:-*,y:-*,Y-* }
QS L_46 { 6-*,c-*,tS-*,k-*,l-*,
m-*,n-*,N-*,p-*,
pf-*,s-*,S-*,x-*,Z-* }
```

Fig. 1. Sample question sets for the left context in a triphone

The resulting question sets were smaller than the hand-written phonetic questions for both German and Spanish. Refer to Table 1 for details. Figure 1 shows a selection of automatic questions produced for German, in SAMPA format.

Many of these questions do cluster sensible phonetic classes of phones together. For example `L_27` clusters diphthongs, `L_24` clusters fricatives with affricates, `L_38` clusters long vowels. Questions such as `L_46`, `L_23` and `L_26` probably wouldn't exist in a phonetically-motivated set as they group phonetically unrelated phones together, however some of these questions could be viewed as composite questions.

4.2. Clustered HMMs

Cross-word triphones were cloned from the twelve mixture monophone systems, down-mixed to one mixture and then state-clustered using either the automatic or manual question sets. In order to enable a fair comparison, we aimed at producing equivalent number of senones in both cases. The final number of senones for each system is shown in Table 1.

The clustered HMM sets were then re-estimated and up-mixed to six mixtures. Table 1 shows the average per-frame log likelihood of each system on the training data. It can be seen that the likelihoods are very similar for the automatic HMMs and the manual HMMs.

4.3. Results

Test sets were constructed from the training corpora. None of the test speakers had been seen in training. A context-free grammar was constructed for each of four tasks: dates, natural numbers, spelling and digits, and off-line recognition results were obtained using the HVite tool. Word error rate results are shown for the final 6-mixture systems in Tables 2 and 3.

Language	Questions	Questions	States	Senones	Log/Phys Models	Avg LogL/Frame
German	Automatic	103	367647	2066	122504/5816	-5.78
German	Manual	116	367647	2001	122504/5256	-5.77
Spanish	Automatic	61	56940	1226	18957/2302	-5.42
Spanish	Manual	106	56940	1250	18957/2913	-5.44

Table 1. HMM Parameters after clustering with automatic and phonetic questions

Corpus Test Set	SpM Dates	SpM Nat num	SpII Letters	SpII Digits
No uttrncs	192	42	1575	363
Manual	15.1	9.0	37.7	4.4
Automatic	17.7	6.2	36.4	5.0

Table 2. WER comparison of 6-mixture, tied-state German HMM systems clustered using manual or automatically derived question sets

Corpus Test Set	VAHA Dates	SpM Dates	VAHA Spell	SpM Spell
No uttrncs	815	2857	750	1731
Manual	6.4	3.7	54.5	38.3
Automatic	6.3	3.5	55.2	37.7

Table 3. WER comparison of 6-mixture, tied-state Spanish HMM systems clustered using manual or automatically derived question sets

Comparing the two German systems, the automatic questions performed better than the hand-written questions on natural numbers and isolated letters; about the same on the isolated digits but significantly worse on dates. The Spanish system trained on automatic questions performed slightly better on three out of four tests.

5. CONCLUSIONS

The results are very encouraging and show that this procedure for training mutual-information based questions could be used as a viable alternative to writing phonetic questions by hand. The automatic question sets do cluster linguistically sensible classes in many cases, and produced smaller question sets that resulted in equivalent or slightly better recognition than the hand-written questions.

The rationale behind writing phonetic questions was to include as many questions as possible that may be relevant, and leave the clustering procedure to select the best question at each point, though this could lead to a larger tree. Questions are generally taken from fields of phonology, acoustic-phonetic or articulatory phonetics of the language, either using in-house linguistic knowledge, or from linguistic literature on the language. The hand-written question sets also contain some composite questions about combinations of features such as `L_BackVowel`. Thus hand-written sets may be larger than necessary and sub-optimal.

The automatic generation of questions is very attractive for a new language where such linguistic knowledge is hard to come by, and even when lingustic knowledge does exist there is uncertainty over how many and which kinds of questions to use. This approach removes such uncertainty and allows the state-clustering process to be fully automated in acoustic model building. The same approach to generating mutual-information based question sets can be utilized in the training of decision-tree-based letter-to-sound rules.

6. REFERENCES

[1] L. R. Bahl et al., "Context dependent modeling of phones in continuous speech using decision trees," in *Proceedings of DARPA Speech and Natural Language Processing Workshop*, Pacific Grove, CA, 1991, pp. 264–270.

[2] Mei-Yuh Hwang, Xuedong Huang, and Fileno Alleva, "Predicting unseen triphones with senones," Tech. Rep. CS-93-139, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1993.

[3] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings ARPA Workshop on Human Language Technology*, Berlin, 1994, pp. 307–312.

[4] John M. Lucassen, *Discovering Phonemic Base Forms Automatically: an Information Theoretic Approach*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, August 1983.

[5] P. deSouza J. Lai P. Brown, V. Della Pietra and R. Mercer, "Class-based n-gram models of natural language," in *Computational Linguistics*, vol. 18, pp. 467–479. 1997.

[6] S. Young, "The HTK hidden Markov model toolkit: design and philosophy," Tech. Rep. TR.153, Department of Engineering, Cambridge University, UK, 1993.

[7] "ELRA catalogue," http://www.icp.grenet.fr/ELRA.

[8] "LDC catalogue," http://www.ldc.upenn.edu.