DOMAIN ADDUCED STATE TYING FOR CROSS-DOMAIN ACOUSTIC MODELLING

R. Singh, B. Raj and R. M. Stern
Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

ABSTRACT

In situations when automatic speech recognition (ASR) systems are rapidly deployed for a new task, the availability of within-domain training data may be limited. In such cases one needs to build the ASR system from other, possibly out-of-domain databases. We refer to the process of building ASR systems for one task domain using data from other domains as cross-domain modelling or CDM. Conventional CDM-based systems perform poorly because the disparity between the triphonetic distributions of the training and test domains is not well accounted for. In this paper we describe two techniques to impose the acoustic-phonetic structure of the task domain on acoustic models built from out-of-domain data. The first technique, called *Extrinsic* CDM, combines decision tree structures obtained from a database close in domain to the task domain with acoustic models that are trained from a third less domain-relevant database. In the second technique, called *Intrinsic* CDM, the task domain data is used to impose the triphonetic distribution of the task domain on the decision trees built from an out-of-domain large database. Both these techniques result in acoustic models which perform better than conventional CDM models.

1. INTRODUCTION

Automatic speech recognition (ASR) systems sometimes need to be trained for a specific task using data that belongs to a different task or domain. This situation may arise, for example, when there are insufficient training data within the specified task domain, such as in rapid deployment situations. We refer to the process of training ASR systems from out-of-domain databases as cross domain modelling (CDM). Historically, CDM has not been a very effective method of building ASR systems for very specialized task domains. It has been known to result in acoustic models which perform poorly compared to models estimated using data from the same task domain. This has been known to happen even when the recording conditions for the training and test databases are similar. Previous attempts at training ASR systems from small amounts of data have included CDM among other techniques (*e.g.* [1]). However, the strategy used in [1] has been to train these systems directly from out-of-domain data, and then to adapt these models to the task domain data. No attempts have been made to account better for the task domain *during* the modelling procedure.

In the absence of good CDM techniques, it has been preferable to always build ASR systems from databases that are close to or within the task domain, even when the available training data are insufficient. Data insufficiency may also arise due to an entirely different reason: overparametrization. Current ASR systems use phones and triphones as the basic units of continuous speech. They model these units statistically for recognition (*e.g.* [2]). Generally, the set of possible triphones for a standard language even within a constrained task domain is very large. This results in a large number of statistical parameters to be estimated which in turn causes data-insufficiency problems. Moreover, even in an ideal situation where the training corpus is very large and there are sufficient data to estimate all these parameters, state-of-art computers cannot store and process all of these parameters without running into serious logistic problems. One is therefore forced to use smaller amounts of data while employing efficient parameter distribution techniques to counter the resultant or anticipated data insufficiency problem.

This reduced set of parameters is obtained by grouping triphones into a statistically estimable number of clusters using decision trees [3]. For ASR systems based on Hidden Markov Models (HMMs), the decision trees result in sharing of output probability distribution functions across states, a procedure well known as *state tying*. Canonically, parameters are distributed (*i.e.* the states are tied) so as to best capture the acoustic-phonetic structure of the training corpus. In CDM, however, the acoustic-phonetic structure of the task domain may be significantly different from that of the training corpus. This disparity alone largely accounts for the degradation in performance when ASR systems are trained from out-of-domain corpora.

In this paper we explore the possibility of compensating for some of this disparity by redistributing the states of HMM-based ASR systems according to the acoustic-phonetic structure of the *task* domain data, rather than that of the training domain data. We focus on conditions of insufficient data, a situation when CDM becomes absolutely necessary, and attempt to improve recognition performance under these conditions. In Section 2 we discuss some of the problems related to CDM. Section 3 describes the strategies we propose to overcome these problems. In Section 4 we describe related experiments and results. In Section 5 we discuss our inferences and present our conclusions.

2. THE PROBLEMS POSED BY CDM

Two databases that cover different task domains within the same language are likely to differ in many ways. Two of these differences which affect acoustic models the most are 1) *Non-overlapping triphonetic coverage*: some of the triphones that are seen in one database may be completely absent from the other, and 2) *Overlapping triphonetic coverage*: some triphones may be common to both databases, but with different relative frequencies of occurrence.

Figure 1 shows the relative frequencies of the occurrences of twelve phones from the CMU phoneset, using the DARPA Broadcast News Corpus [4]. It also shows the frequencies of occurrence of the same phones in the CMU Communicator Database, which is described later in this paper. The figure brings out the differences in relative frequencies rather clearly.

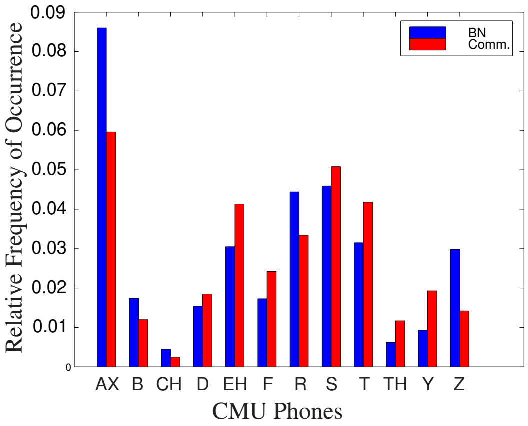

Figure 1: Comparison of the relative frequency of occurrence of twelve phones in the CMU phoneset. In each pair, the bar on the left denotes the BN database and the bar on the right denotes the CMU-Communicator database.

The relative frequencies of occurrence for the phones differ for almost all the phones shown, and this is a representative sample from the CMU phoneset. At the triphonetic level this disparity of relative frequencies is even greater across task domain databases.

An additional consideration for ASR systems is that it may be more important for some triphones to be distinguishable from one another in one database, than in the other. As a result, it would be advantageous to model these triphones separately for one task, while this may result in no gain in the other task.

There are therefore two differences to be considered: (a) the basic difference in the acoustic-phonetic nature of the databases, as specified in points 1 and 2 above, and (b) the manner in which statistical parameters are distributed during training. The distribution of statistical parameters is mediated by decision tree structures which partition the training data into logical classes based on some objective function. These decision trees perform the additional role in ASR systems of associating triphones not seen during training with one of these logical classes, thereby permitting the modelling of these "unseen" triphones during recognition. The structure of the decision trees controls (b), and thereby also controls triphone distinguishability. If these trees are built entirely from one database, their structure would conform to the triphonetic distribution of that database. This structure may extend to triphone classes which are not present in the task domain (non-overlapping regions) causing entirely superfluous models to be built.

While (a) cannot be amended, (b) may be modified by using state tying procedures to alter the triphonetic distributions given by the decision trees to be more representative of the task domain database. In the following section we address the problem of achieving domain specificity in CDM by appropriately-motivated state tying procedures which ensure state clustering patterns that are relevant to the task domain.

3. DOMAIN INFERENCE BY MODIFIED STATE TYING

In HMM-based ASR systems, decision trees are built by recursively partitioning the triphones associated with each node of the tree so as to maximize the likelihoods of the vectors belonging to the triphones in each of the partitions. These likelihoods are computed on the distributions inferred from the vectors within the partition. Therefore, each partitioning results in an increase in the overall likelihood. A separate decision tree is built for each phone. The root of the decision tree includes all the triphones associated with that phone. In a complete tree, each leaf represents a single triphone. Parameter reduction is achieved by pruning the leaves of the tree progressively to eliminate leaf and node pairs that resulted in the lowest increase in likelihood. The resulting pruned tree has leaves which represent groups of triphones rather than single triphones, and that have the maximum likelihood possible given the structure and training corpus used to produce that particular decision tree. Each leaf in the pruned tree is a tied state and the resultant distribution of states models the acoustic-phonetic properties of the corpus that was used to build and prune the trees.

However, it is still necessary for the distribution of states to represent the task domain. Ideally, this could be achieved by building the decision trees from the task-domain data. Since we begin by assuming that the data are insufficient, we cannot get good estimates of the distributions used to generate decision trees using only data from within the task domain. Moreover, trees generated using only task-domain data are likely to have nodes or leaves corresponding to non-overlapping triphones for which training data would not be available.

3.1. Extrinsic Domain Inference

One way to bypass the data-insufficiency problem is to use data from yet another task domain that is similar to the current task to generate the decision trees. The acoustic (or recording) conditions of the new data are irrelevant. These decision trees can then be used to train the ASR system using the larger cross-domain training corpus. We refer to this process as *Extrinsic Cross-Domain*

Modelling, since the state tying is achieved through decision trees which are not directly provided by either the main training corpus or the task domain.

Unfortunately, Extrinsic CDM requires the availability of a third database covering a similar task domain. Such a database may not always be available, and therefore we do not recommend Extrinsic CDM as a method of choice in situations where Intrinsic CDM (described below) is feasible.

3.2. Intrinsic Domain Inference

A second method to achieve CDM, referred to as *Intrinsic* CDM uses the out-of-domain training corpus to generate the decision tree structure, and uses the *task* domain corpus to prune the decision tree to obtain the desired distribution of parameters. Here we take advantage of the fact that *while a great deal of data is required to estimate distributions, only a small amount of data is needed to validate them.* (For example, recognition is performed on very small amounts of data).

The data corresponding to the various triphones of each phone are used to compute the likelihoods of the each of the nodes in the decision tree belonging to the phone. The data used to compute the likelihoods at each node are the vectors from the triphones in the task domain associated with that node. The distributions from which the likelihoods are computed are obtained from the training corpus. Because of the nature of the decision tree building process, all triphones in the task domain get associated with at least one of the nodes in the decision tree, including those triphones that were never seen in the corpus used to build the tree [3]. As a result, the likelihoods computed at each node in the decision tree represent the acoustic-phonetic properties of the entire task domain corpus. The decision trees are then pruned so as to maximize the likelihood of the leaves of the pruned decision tree by the standard procedure of sequentially eliminating the nodes with the lowest increase in likelihood.

Nodes in the decision tree that represent triphones that do not occur in the task domain corpus have a zero likelihood, and also result in zero increase in likelihood and thus get pruned out. As a result, no modelling effort is directed at events that are not likely to be observed in the task domain corpus. As a result, the pruned decision trees are also less likely to group triphonetic units together that are best kept separate for the task domain. When the test domain is the same as the domain from which the training data were derived, Intrinsic CDM reduces to the conventional state-tying procedure.

Both Intrinsic CDM and Extrinsic CDM result in distributions of parameters that are not optimal for the training corpus itself. As a matter of fact, some of the resultant tied states may have insufficient data in the training corpus for proper estimation of distributions. A greater improvement in performance can therefore be expected posterior to CDM by adapting the Intrinsic CDM-based acoustic models to the task domain data in these cases. In adaptation we use the task domain data to perform Maximum Likelihood Linear Regression (MLLR), Maximum *a-posteriori* (MAP) or other smoothing operations on the probability distributions of the HMMs. This adaptation must not be confused with the CDM techniques themselves which only determine *how* the parameters are distributed amongst the various phonetic units, and not the *actual values* of the parameters.

4. EXPERIMENTAL RESULTS

The CMU Communicator is a conversational system that works with users in a travel planning domain. Users attempt to interact with the Communicator as they would with a real travel agent for making real reservations for travel within the United States. The task covers topics such as airplane, hotel, and car reservations and inquiries. The data includes noise conditions typical of any telephone conversation. ASR systems were built for this domain using the Extrinsic and Intrinsic cross-domain modelling techniques described in the previous section.

For Extrinsic CDM, decision trees were generated using data from the Air Travel Information System (ATIS) database, which consists of about 12 hours of broadband speech. Utterances in this database involve inquiries about airline schedules and travel and is therefore a subset of the communicator domain. The data are clean broadband speech and do not represent the recording conditions or the noise conditions of the Communicator. The acoustic models themselves were trained with 72 hours of Broadcast News (BN) speech that had been filtered to telephone bandwidth. Decision trees were built and pruned using the ATIS database and the tied states were trained with the BN database.

For Intrinsic CDM, training used only the Broadcast News data and the Communicator data. The Broadcast News corpus is quite large and has a wide coverage of topics, so it is expected to have a comprehensive triphonetic coverage. Noise conditions within this corpus are also varied and may cover some of the noise conditions present in the Communicator data, (*e.g.* background chatter and reverberation). The decision trees were built using the Broadcast News corpus and were pruned using the Communicator data. The Communicator data that was used here and for adaptation of the probability distributions of the HMMs consisted of about 2.5 hours of recorded transactions. Figure 2 shows the resultant distribution of tied states resulting from each of the schemes described above. Each bar represents the total number of tied states used to model a particular phone. This would correspond to the total number of leaves in the pruned trees for that phone.

Comparison of Figure 2 with Figure 1 reveals an underlying similarity. By comparing the BN portions of the two figures it can be seen that the optimal number of tied states used to model any particular phone is somewhat monotonically related to the relative frequency of occurrence of that phone. Also, while the Broadcast News portions of Figure 2 do not resemble the relative distribution of phones in the Communicator database in Figure 1, the distribution of tied states obtained by using Intrinsic CDM

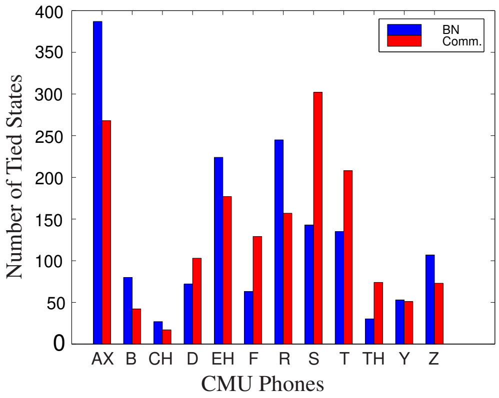

Figure 2: Comparison of the number of tied states obtained after pruning decision trees using normal pruning and Intrinsic CDM for the phones indicated. In each pair of bars the left bar represents the number of tied states obtained after conventional pruning based on training data based pruning and the bar on the right represents the number obtained after pruning based on Intrinsic CDM using the Communicator data

Method	Word Error Rates (%)	
	unadapted	adapted
Training with BN	48	45
Training with Comm.	53	–
Extrinsic CDM	45	42
Intrinsic CDM	45	40

Table 1: Word error rates for comparison of acoustic models built using regular and cross domain strategies

matches the relative distribution of the phones in the Communicator database. This clearly indicates the effectiveness of Intrinsic CDM in capturing the acoustic-phonetic nature of the task.

The test set consisted of 700 utterances of particularly noisy Communicator data. Untrained users were asked to call the Communicator from a crowded auditorium and the resulting recordings consist of a high level of background noise and disfluencies. The experiments were run using the Sphinx-III semi-continuous system using 5000 tied states. Adaptation to test domain data was performed by interpolation of the prior-probabilities of the HMM with those obtained by training using test domain data.

Table 1 shows the recognition results corresponding to models built using extrinsic and intrinsic CDM techniques, with and without adaptation to the test domain. We note that greater improvements are obtained on the adapted models than on the unadapted models, especially in the case of Intrinsic CDM. This is understandable given the fact that the decision trees in Intrinsic CDM are pruned to maximize the likelihoods of the *test* domain data. Since adaptation also attempts to maximize the likelihood of the test domain data, the combination of Intrinsic CDM trees and the adaptation is most effective in modelling the test domain data.

5. DISCUSSION AND CONCLUSIONS

One could argue that with the current communication technology and its fast-rising affordability, CDM may never be necessary as data insufficiency may never arise. However, we anticipate data-insufficiency problems to crop up in situations of rapid deployment – in systems which attempt to rapidly learn to recognize within a very specific task domain, in possibly a very new language or its dialectic variations. CDM may also be called for when the acoustic conditions of the test data differ significantly from the available within domain training data.

From our experiments we conclude that both Extrinsic and Intrinsic CDM can be used to enhance recognition accuracy. Figures 1 and 2 show that there is a tentative relationship between the optimal distribution of parameters and the relative frequencies of occurrence of the phonetic units. This indicates that in some instances triphone counts empirically derived from the task domain can be used for pruning the decision trees.

The CDM technique proposed *cannot* be used where the objective function used in the decision trees is not verifiable, or does not permit reliable statistical conclusions to be drawn from small amounts of task domain data. Also, likelihoods are usually computed on every frame and are proportional to the durations of triphones. Computing per frame likelihoods may yield more balanced decision trees. In this case the recognizer would also have to use normalized likelihoods as the recognition criterion.

6. ACKNOWLEDGEMENTS

This research was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

7. REFERENCES

[1] Siu, M., Jonas, M., Gish, H., "Using a Large Vocabulary Continuous Speech Recognizer for a Constrained Domain with Limited Training", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, May 15-19, 1999, Phoenix, Arizona.

[2] Lee, K., Hon, H., and Reddy, R., *An overview of the SPHINX speech recognition system*, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38, pp.35-45, 1990.

[3] Hwang, M. Y., "Sub-phonetic Acoustic Modelling for Speaker-Independent Continuous Speech Recognition", PhD Thesis and Computer Science Tech. Rep. CMU-CS-93-230, Carnegie Mellon University, 1993.

[4] Graff, D., "The 1996 Broadcast News Speech and Language-Model Corpus", Proceedings of the 1997 DARPA Speech Recognition Workshop, Chantilly, Virginia, 1997.