MULTIPLE TASK-DOMAIN ACOUSTIC MODELS

Andrej Ljolje

AT&T Labs - Research
180 Park Ave., Florham Park, NJ 07932, USA
alj@research.att.com

ABSTRACT

Many speech recognition applications require the recognizer to perform at peak recognition accuracy across many different domains. Examples of different domains are general English, digits, names, alphabet, etc. Here we show a way to preserve the simplicity of a single acoustic model while providing domain specific recognition speed and accuracy. This is achieved by employing an extended phoneme set that keeps a subset of phonemes specifically for a particular domain, and a context dependency specification that allows cross-word, cross-domain phonetic context dependencies. Testing on a names recognition task going from a wrong domain (general English) model to a multiple domain model (general English, alphabet, names) the error rate is reduced by more than 50%. Domain-specific model trained only on the names data further reduces the error rate by more than 50%.

1. INTRODUCTION

Traditional development of acoustic models for speech recognition centered around collecting large amounts of training data and test data, training the acoustic model on the training data and testing on the test data. Both the training and test data came from the consistent pools of speakers, recording utterances that are consistent in recording conditions, topics and language structures. We consider those as in-domain or domain-specific acoustic models and recognition results. It has always been known that any changes to any of the domain characteristics can drastically reduce recognition performance. Examples of the types of differences include: changing the microphone from a close-talking high quality microphone to a telephone receiver; training on continuous speech, testing on isolated utterances; training on general English, testing on small vocabulary tasks like digits or alphabet.

Recent proliferation of dialog management systems have imposed many potentially debilitating requirements on acoustic models to perform under many different conditions [3]. In an interaction between human users and an automated speech recognition system guided by an automatic dialog manager, the speech recognition system would be expected to handle different types of utterances. For example, users might provide a request or a description using general English. Users might provide account numbers that contain digits and spelled letters. They might provide telephone numbers as digit sequences, their names, generic responses to confirmation requests, dates and times, addresses, etc. Ideally, each one of those types of utterances, which we here call domains, would be handled by the domain-specific model that was trained using in-domain training data and using in-domain trained language model. The problem with that approach is that it requires training many different models with carefully filtered training data which would need to be swapped many times during the interaction between users and dialog systems. Also, it would create unsurmountable problems if a single utterance contained speech from more than one domain. An obvious example is of a digit string, like a telephone number, being embedded inside a general English context.

Traditionally, this kind of a problem was infrequent, and it was handled by pooling the data from all the domains into a single sub-word unit based acoustic model, typically using clustered triphones as acoustic units. We describe an approach that allows usage of a single acoustic model that provides recognition performance equivalent to a model trained using in-domain training data and tested on in-domain data, including the difficult case when data from multiple domains occurs in a single contiguous utterance.

2. MULTIPLE-DOMAIN MODEL STRUCTURE

The underlying structure of all of the acoustic models described here is based on triphonic HMMs with three states left-to-right HMMs per triphone. They all use four silence models, two with single state and two with three state left-to-right HMMs. All the states were modeled with a 10-component Gaussian mixture except for one single state silence state which was modeled with a 24-component Gaussian mixture.

Given the basic structure of the acoustic model the only flexibility that could allow modifications that would result in multiple domain-specific acoustic models within a single model structure was the selection of the phoneme inventory. An extended phoneme set was selected so that different groups of phonemes represent different domains. This resulted in domain-specific sub-models. Representation of such an extended phoneme set is best made by using the conventional phoneme representation and reserving it for the general English domain. All other domains use the same phoneme set (or its subset) with a suffix (a diacritic). However, the phoneme selection is not limited to the conventional English phoneme set, and an example of such divergence is in the selection of a head-body-tail subword representation for the digits. An example of a small dictionary representing a typical small subset of words from the general English, names, digits and alphabet domains can be seen in Table 1. Here we use the following suffixes as diacritics: “_d” - digits; “_a” - alphabet; “_n” - names.

Given the extent that the subword inventory in the example here differs from a conventional phoneme set, it should be referred to as a subword set instead. However, due to the possible confusion and an ingrained custom to use the term phoneme set regardless of the subword units’ characteristics, the rest of the paper will continue to use the common term phoneme set.

The acoustic models in the experiments described below all

WORD	BASEFORM
am	ae m
I	ay
therefore	dh eh r f ao r
think	th ih ng k
A	ey_a
J	jh_a ey_a
L	eh_a l_a
Joe	jh_n ow_n
Smith	s_n m_n ih_n th_n
0z	h0_d b0_d t0_d
0o	ho_d bo_d to_d
1	h1_d b1_d t1_d

Table 1. A small dictionary showing different phoneme sub-sets for general English, alphabet, names and digits, with digits represented using a non-phonemic subword units (head-body-tail).

had an extended phoneme set as shown in Table 1. Also, the model structure was based on three-state left-to-right context-dependent HMMs. One difficulty with such an approach is that the training data is often, as it was here, composed of domain-specific collection sets. Together different sets can cover several domains, but databases that have a good representation of several domains is rare, especially if utterances containing speech from several domains is required. With the phoneme set used here, it was impossible to find any database that provided even a small part of possible context across domains. Data across contexts within a single domain is often sparse which is why we resort to tree-based context clustering to avoid the data sparsity problem. This problem is drastically more pronounced across different domains. This problem can be resolved by ignoring cross-word context, and thus ignoring cross-domain contexts as was done in [1]. We resolve this problem by defining the subword units in two different ways for two different purposes. They are defined as phonemes with diacritics for the purposes of dictionary building and recognition search. However, we define them in terms of their phonetic features for the purposes of context definition. Thus different phonemes “”ae” and “ae_n” can be considered equivalent phonetic contexts, even though they are considered different entries in the dictionary or the search network [2]. In addition, during the tree building process to define state tying across different contexts, which is driven by the amount of training data and differences in the data across different contexts, questions can be asked about which domain the phoneme belongs to in addition to the questions about the phoneme phonetic features. This would allow separation across domains for contexts if sufficient amounts of cross-domain training data speech are available and the phonemes are different in those contexts. In all the experiments described below, this feature was not used and only phonetic feature-based questions were used in building the context-dependency trees.

3. TRAINING AND TEST DATA

The training data was telephone speech collected in more than twenty different internally consistent sets. Most of the datasets contained data from only one domain, although some contained data from two domains. Very rarely data from more than two domains were included in the same utterance. The most common combination of domains was alphabet + digits and general English + digits. The total amounts of training data for each domain is shown in Table 2.

DOMAIN	# WORDS
general English	1,635,171
digits	221,031
alphabet	100,926
names/isolated utt.	73,301

Table 2. The amount of training data varied by more than an order of magnitude between different domains.

DOMAIN	# STRINGS	# WORDS
alphabet/digits	996	2110/4859
first/last names	551	1107

Table 3. The amount of test data for the two domains used in the experiments. The alphabet data is less than a third of the test set for the alphabet/digits domain.

Two test sets were selected to demonstrate the effects of domain-specificity in acoustic models. The first set consists of isolated utterances of peoples’ names. Each utterance consists of spoken first and last names. The second test set consists of recordings of seven character account strings which consist of letters and digits. A significant number of strings contained speech extraneous to the domain, which was not included in the language model for the task (eg. “6 4 R as in Raymond N as in Nancy 0z 7 0z”). The language models in both cases only considered in-domain strings as valid, and enforced in-domain recognition output. The vocabulary size for the names task was 2000 words, and for the alphabet/digits task it was the 25 letters and 11 digits, including both “zero” and “oh” versions of digit “0”. The letter “O” was not used to avoid the obvious confusion between the letter and digit versions both of which sound the same: “oh”. The amount of available test data is shown in Table 3.

4. RECOGNITION EXPERIMENTS

Two multiple-domain acoustic models were built in order to perform the comparison between models with different levels of domain dependency. Both models were of exactly the same size and structure shown in Table 4.

The combined size of all the domain-specific parts of the model and the distributions modeling silences was 7895 states and 22815 HMMs.

The first acoustic model implicitly only used domain-specific data to train all of the model parameters, as defined in the dictionary specifications. In this case all the states of all of the HMMs corresponding to the alphabet domain phoneme set were trained using all of the occurrences of spelled letters in all of the training data sets. None of the spelled letters’ data was used in training

DOMAIN	# STATES	# HMMS
general English	5235	16329
digits	660	495
alphabet	704	1891
names/isolated utt.	1296	4100

Table 4. The number of states and HMMs varied across domains depending on the amount of the available training data as well as recognition speed and accuracy.

any of the remaining states of the model. Similar differentiation applied to every distinct domain included in the model: general English, digits, alphabet and names. This model was used to test the accuracy of the domain-specific model when tested on domain specific data. We do that by testing alphabet and digits models on combined letters and digits test set. The same model was used to see how bad the performance is when the general English model, which is trained only on general English data, is tested on a different domain. We use the same, now out of domain test set as before, names and letters/digits.

The second model is an exact duplicate of the first model for the digits, alphabet and names domains. However, the general English part of the model was trained on the data set which was the combined set of general English data, alphabet data and names data. It is referred to as the multiple domain model, as it corresponds to the more traditional approach to acoustic modeling where all the data was combined to build a single monolithic acoustic model. Access to the different parts of the acoustic model in the recognition process is controlled by the use of domain specific diacritics in the test setup dictionary. All of the recognition performance plots below have three sets of results.

The first corresponds to matched conditions of using domain-specific part of the model to recognize speech from the same domain (domain-specific). The second corresponds to matching one of the combined domains in the model trained using data from several domains and the test domain (multiple domains). The third corresponds to a mismatched model and test data domains, when data from one domain is used to train a part of the model, in this case general English, and is tested on different domains, like alphabet and names (wrong domain).

All the experiments were performed with endpointing turned on and off using *Watson 4.1* recognition platform. It is an AT&T platform which is used for AT&T deployed speech recognition applications.

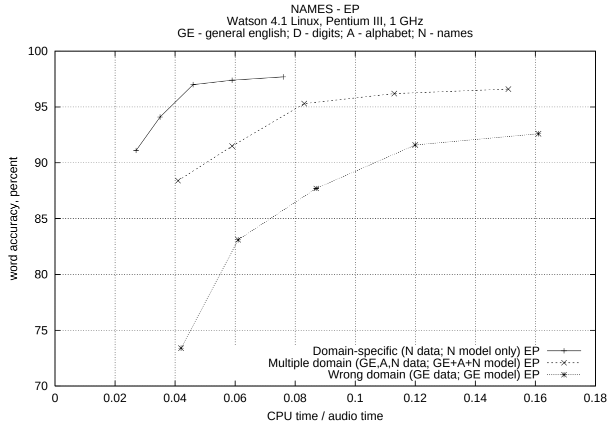

Fig. 1. Comparison of word recognition performance on the names recognition task across different levels of domain-specificity, using endpointing

Figure 1 shows the results on the names test set, with the best performance with matching training and test domains, less accurate and slower results with the multiple domain model, which is significantly larger than the names domain model, and the worst performance with the wrong domain model. Similar performance is achieved when endpointing is not used as shown in Figure 2.

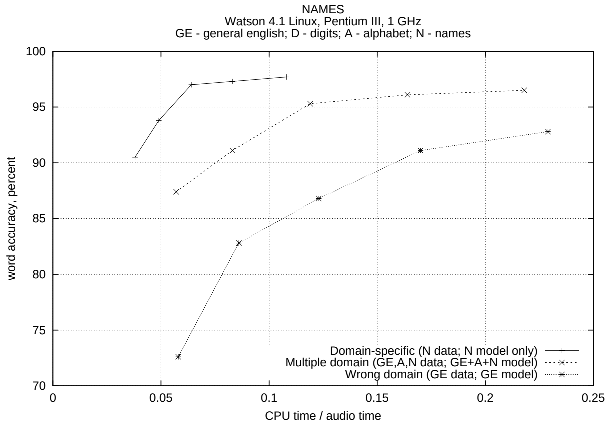

Fig. 2. Comparison of word recognition performance on the names recognition task across different levels of domain-specificity, without endpointing

The difference in recognition speed is significant because while the decoder is in the initial silence it has to hypothesize and score all the possible states following the silence. In the case of names recognition there is a very large number of possibilities (large branching factor), which keeps it slow. Once the first few phonemes have been recognized the number of options is very small, at a reasonable beamwidth, and the rest of the recognition takes very little time. Removing the silence speeds up the average recognition time.

The experimental setup is not as simple or as clear in the case of the second test set. Ideally a test set would have been available consisting of only spelled letters. The nearest approximation which was available was a test set consisting of account numbers which contained strings of seven letters or digits. There were more than twice as many digits as letters. In all the experiments the same digits in-domain model is used and the performance is quoted for the whole test set, combining digits and alphabet results. The language model enforced a 7-character string length, minimizing the effect of having digits in the string together with letters. The word accuracy performance is shown in Figure 3 and string accuracy is shown in Figure 4.

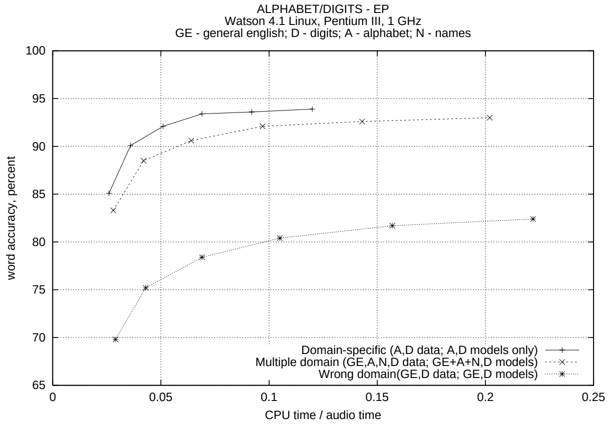

Fig. 3. Comparison of word recognition performance on the names recognition task across different levels of domain-specificity, using endpointing

The performance difference between the three levels of domain-

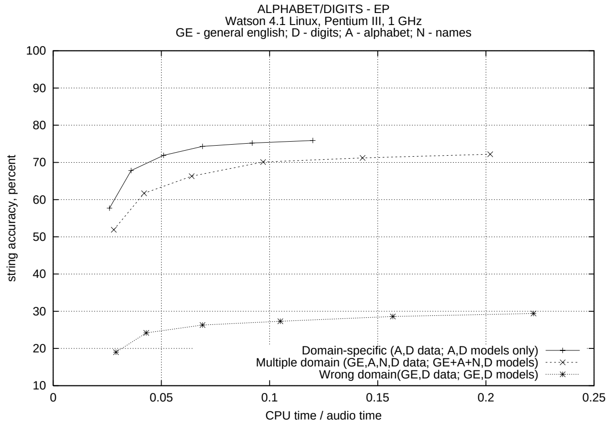

Fig. 4. Comparison of word recognition performance on the names recognition task across different levels of domain-specificity, using endpointing

specificity is slightly different than for the names task. However, there is the same pattern of significant recognition error reduction as model becomes more domain specific. The recognition speed is almost the same on this task as the names task when endpointing is used, due to the small amount of silence included in the speech files, and the reduced branching factor while searching through the utterance-initial silence. The marginally slower speed of recognition, but with similar accuracy, can be seen in word accuracy and sentence accuracy plots shown in Figures 5 and 6, respectively.

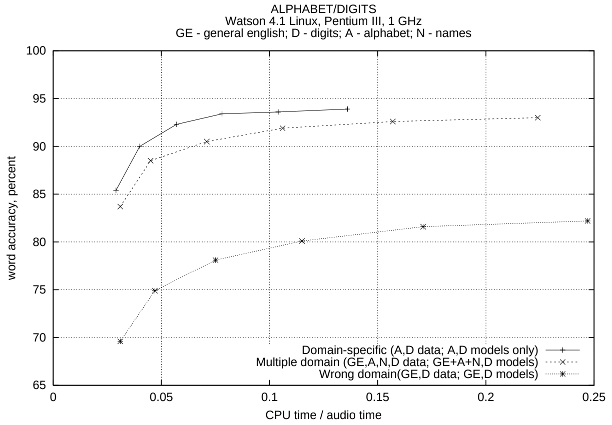

Fig. 5. Comparison of word recognition performance on the names recognition task across different levels of domain-specificity, using endpointing

5. CONCLUSIONS

Comparison of recognition performance between three acoustic models of different levels of domain-specificity clearly demonstrates both the accuracy and speed advantages of matching training and test data characteristics. Here we showed a simple approach for combining the simplicity of a single acoustic model implementation with domain-specific performance using extended phoneme sets and phonetic feature-based context dependency clustering. A model built using the proposed structure achieves domain specific performance, even if more than one domain is present in a an utterance, without any changes to the decoder.

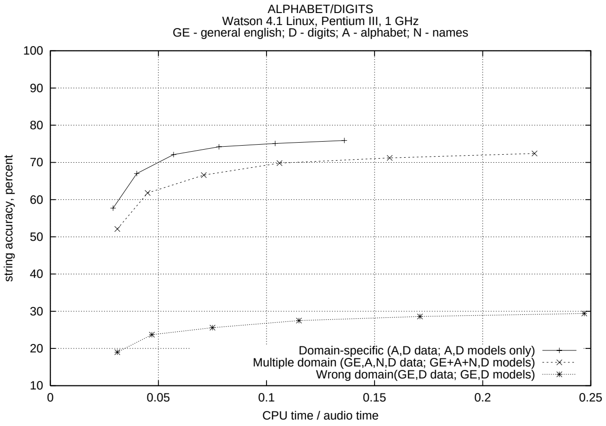

Fig. 6. Comparison of word recognition performance on the names recognition task across different levels of domain-specificity, using endpointing

6. ACKNOWLEDGMENTS

The author gratefully acknowledges help with the training database from Anne Kirkland, various model conversion and recognition setup issues from Vincent Goffin and generating recognition performance within the Watson 4.1 recognition system by Joe d'Andrea.

7. REFERENCES

[1] S. Deligne, E. Eide, R. Gopinath, D. Kanefsky, B. Maison, P. Olsen, H. Printz and J. Sedivy, "Low-Resource Speech Recognition of 500-Word Vocabularies," In *Proceedings EUROSPEECH*, 2001.

[2] M. Riley, "Tree-based models of speech and language," Proc. of *Interface* 1991, Seattle, WA.

[3] M. Walker, A. Rudnicky, J. Aberdeen , E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, R. Prasad, S. Roukos, G. Sanders, S. Seneff and D. Stallard, "DARPA Communicator Evaluation: Progress from 2000 to 2001," In *Proceedings ICSLP*, 2002.