A CTS TASK FOR MEANINGFUL FAST-TURNAROUND EXPERIMENTS

B. Chen‡, Ö. Çetin‡†, G. Doddington‡, N. Morgan‡, M. Ostendorf†, T. Shinozaki†, and Q. Zhu‡

†Department of Electrical Engineering, University of Washington, Seattle, WA

‡International Computer Science Institute, Berkeley, CA

Contacts: morgan@icsi.berkeley.edu, mo@ee.washington.edu

ABSTRACT

This paper describes a small task paradigm used by researchers in developing new front end and acoustic modeling techniques for conversational speech recognition. The task definition involves both test data selection to support a small vocabulary and training data selection to reduce training costs. It was designed to support rapid experimentation and still give results that translate to conversational telephone speech recognition tasks using state-of-the-art systems. We discuss experiments in scalability and provide details of a recent small task definition and performance of baseline systems on this task. We also include results from experiments in training data selection, showing some benefit to sampling methods that de-emphasize outliers.

1. INTRODUCTION

A challenge in developing radically new approaches to speech recognition is the high cost of experimentation for large vocabulary tasks, both because of impractical decoding costs for new types of acoustic models (e.g. non-HMM) and/or the excessive training costs associated with the large volumes of training data being used by state-of-the-art systems. However, there is justifiable skepticism of results reported on small vocabulary or isolated word tasks; it is frequently the case that gains on such tasks do not translate to larger tasks and state-of-the-art systems.

In order to balance experimentation time, which is critical to minimize for exploring a large space of new techniques, with the need for scalability, we developed a small vocabulary conversational speech task with reduced training and test set definitions that were based on the same data as used by EARS RT systems. The availability of a small task was especially critical in developing MLP-based novel features, as well as for our work in multi-rate hidden Markov modeling aimed at better using long term features. The strategy was to use a small system to test new features with fast turn-around time. This allowed us to try many different things at an early stage, and later use a larger task to get more convincing results that a new development really gives good performance. By using a small task that is tightly coupled to the target task, we found the improvement in a small system does often translate to large systems. For example, we observed that improved feature extraction techniques that led to a 10.5% relative word error rate (WER) reduction on a 500 word conversational telephone speech (CTS) task [1] translated to a 9.9% relative gain on the full task when integrated into the complete SRI evaluation system.

Encouraged by the success of this work, and needing to change the training/test scenario to better track changes in current CTS speech collections, we have developed a new "small task" that is described in this paper. In addition, we have developed a baseline system with the publicly available HTK software package. Our hope is that this resource will allow a larger number of researchers to do meaningful speech recognition experiments on a CTS task.

There are several advantages to using this intermediate task before scaling up to a full CTS system. First, the recognition vocabulary was reduced by choosing the most common words from a standard CTS corpus, so it is likely that the error rate reduction would apply to the larger task as well. Second, there are many examples of the common words in the training data, so less training data is to some degree reflective of the situation for less frequent words in a larger training set. The smaller training set speeds up training time linearly for HMMs, but almost quadratically for neural networks. Finally decoding complexity in this task is smaller, which reduces experiment turnaround time without relying on an N-best rescoring paradigm, which has the disadvantage of the "Rover effect" obscuring the true benefits of modeling innovations.

In the remainder of the paper, we provide more details on the small task definition and experiments described above that led to scalable performance results. Next, we introduce the new task definition and describe baseline systems built for the new task. Finally, we describe additional experiments in training data selection and discuss directions for future research.

2. EXPERIMENTS IN SCALABILITY

Our initial efforts in experimenting with a small task involved recognition of Switchboard data using a vocabulary of the 500 most commons words in that data. The test data represented a 1.4 hour subset of the standard NIST 2001 CTS evaluation test set, with utterances chosen so that the top 500 most frequent words covered the vocabulary in each utterance with less than 10% out-of-vocabulary (OOV) rate. Since many utterances had a lower OOV rate, the overall OOV rate of the test corpus was 3.2%. (For reference, the 38k vocabulary SRI system has an OOV rate of roughly 1% on this data.) An independent 1 hour development set was selected from the same original test set in the same way, for use in tuning system parameters like word transition weight and language model scaling. The training set of the 500-word CTS task contains 23 hours of randomly selected speech mainly from the Switchboard corpus.

A standard HMM system based on the SRI framework was trained on this data, to explore the integration of MLP augmented features with a PLP baseline. The lexicon incorporated multi-word versions of the 500 words in order to better model pronunciation variability, so the actual decoding lexicon was quite large, with 1347 multi-words and an average of 12 pronunciations per entry. The language model used in both the 500 word task as well as the full vocabulary task was the first-pass bigram language model used by SRI for the large vocabulary evaluations in 2000. On this task, we obtained a 10.5% WER reduction from 43.8% with PLP baseline to 39.2% with the MLP augmented feature [1].

At ICSI, we further use the full NIST 2001 evaluation set of 6.3 hours as the final test-bed to see the performance of the feature on a full CTS task. The training set for this task contains 64 hours of speech from the Switchboard corpus. We got 10% WER reduction from 39.1% with PLP feature to 35.2% with the MLP augmented feature. This experimental setup is also used to test the translation of the gain developed with a simple baseline to more complicated baselines [2]. Better baseline features include HLDA on PLP features with three derivatives (HLDA-PLP). With the HLDA-PLP baseline, we got a 8.6% WER reduction from 37.2% to 34.0% with the MLP augmented feature. When further adaptation is used, we got a 8.9% WER reduction from 35.8% to 32.6%.

The idea of developing features with a small task and further testing on a larger task worked well in the past two years. Further tests on features are done at SRI with full Switchboard training of 400 hours, where researchers at SRI also test the feature with a more complicated system including MMIE training, rescoring, and ROVER. In these tests, the features developed by the NA team provided WER reductions ranging from 6% to 8.2% [2]. More recently, the full SRI evaluation system, incorporating over 2000 hours of training data (including both Switchboard and Fisher data) and MPE training also made use of these features. The improvement from the new features was a 9.9% relative WER reduction on both the 2004 development set (from a baseline of 17.2%) and the 2004 evaluation set (from a baseline of 20.3%).

At the same time as ICSI was exploring scalability, UW was exploring implementation within the HTK [3] and GMTK [4] frameworks. For these systems, the publicly available decoders are not as highly tuned as in state-of-the-art decoders such as the SRI system. In this case, it is not practical to decode with the large number of pronunciation and multi-word alternatives in the SRI dictionary. Given the success that others have had with single pronunciation lexicons [5], we explored different WER operating points by using somewhat larger vocabularies. We constructed a new test set by selecting utterances that have an utterance-level OOV rate less than 7.5% according to 500 most frequent words. With a 2500-word vocabulary, the OOV rate on this set was 1%, which is similar to the large vocabulary system. The results of these experiments influenced the direction taken in updating our small task definition for the next stage of research.

3. 2005 NA TASK DEFINITION

While the CTS-500 task proved to be very useful, it was based entirely on Switchboard data and an old test set. In order to keep pace with developments in the EARS task definition, we decided to define a new task that incorporated both Fisher and Switchboard data, and a new test set based on the RT03 evaluation set. The details of the new task definition follow.

Test sets. As in the previous task definition, test segments were selected based on the percent of OOV word tokens in the segment. Selected segments were limited to those segments with an OOV rate of no greater than 10%. (Thus, all selected segments with 10 or fewer words contain no OOV words.) Word pre-processing, performed prior to segment selection, included the following two steps: 1) words that contained underscore characters (e.g. acronyms, compound words) were split into their constituent parts, and 2) "(%HESITATION)" was mapped to "uh". As an alternative selection option, non-word tokens (i.e. "reject" words, laughter and disfluencies) were excluded from the count of OOV words prior to computing the segment OOV rate. A total of 2.5 hours of data was selected from the RT03 evaluation test set. We then divided the data selected into a 1.2 hour tuning set and a 1.3 hour test set. The ratio of utterances from the Fisher corpus relative to those from the Switchboard corpus (1:1.3) is similar in both the tuning and test sets.

Lexicon. The overall size of our 500-word vocabulary dictionary was 4.5k, since it included 1.5k multi-words and average of 2.3 pronunciations for each word/multi-word. The OOV rate for the new test set was 3.5%. Since the OOV rate was relatively high and because of our experience trading off OOVs for multi-word entries, we considered adding words to the vocabulary. We added 500 words with the highest frequency following the original top 500 words. For the newly defined 1000-word vocabulary, we repeated the same process as the original dictionary to expand to multi-words and to prune the pronunciations. Although the vocabulary size was doubled, the increase of the dictionary size was only 13%. This was because the multi-words were made only for highly frequent word sequences. Therefore, there were not many cases where the added words formed multi-words with the top 500 words. Compared to the 500-word vocabulary case, the OOV rate and word error rate were significantly decreased from 3.5% to 2.3% and from 44.1% to 42.0%, respectively. Since the increase of the dictionary size was small, the recognition time using HTK was about the same as for the 500-word vocabulary system.

Initial training data selection. A new training set was selected consisting of approximately 32 hours of speech (16 hours for each gender) coming from a mixture of Fisher (native speakers only) and Switchboard training utterances. Training data was selected by uniformly sampling the Switchboard and Fisher training sets, with the constraint that the two sources would comprise roughly 40% and 60% of each 16 hour subset, respectively. We use this 32 hour training set for training our HMMs, but for neural net training, we further divide our two new 16 hour training sets into 14.6 hour training sets and 1.6 hour cross-validation sets.

Language model. The language model (LM) was made by projecting the 2004 CTS evaluation LM onto the 1000 vocabulary. Any words that were not in the vocabulary were mapped to a special word "reject". Because of this procedure, the resulting language model had very high unigram probability and a large back-off weight for the "reject" word. We found this caused a problem for some decoders that approximate the bigram probability $P(word_2|word_1)$ as the maximum of bigram and unigram probabilities $\max\{P(word_2|word_1),\ P(word_2)\}$. Those decoders that adopt the approximation for an efficient search network representation include HTK, which is one of the most widely used decoders. To support these decoders, we also made an LM in which the "reject" word was removed from the vocabulary. The probabilities of the LM were renormalized after removing the "reject" word.

4. BASELINE SYSTEMS

We have developed two HMM-based systems for this task: one based on the SRI Decipher infrastructure, and a second using publicly available HTK software. Both are simplified systems in that they use only ML (not MMIE or MPE) training. Both systems also use the same baseline features: PLP with the first two derivatives, computed with vocal tract normalization and mean and variance normalization.

In the system based on the SRI infrastructure, gender-dependent models are trained through 7 iterations (2 on context-independent models, 2 on context-dependent models, and 3 on genone-clustered models) [6]. The trained model has 1498 Gaussian clusters for female and 1725 clusters for male. Each Gaussian cluster contains 64 Gaussians. The dictionary with multi-words and multiple pronunciations is used in a first pass of decoding with a bigram LM. This decoding generates an N-best list, which are rescored with optimized language model weight and word transition weight. We use the tuning set to determine the best LM weight and the word transition weight, then apply the new weights in the N-best list rescore to get the final recognition result on the 1.3 hour test set. With this setup, the baseline PLP feature achieves 41.7% WER.

We also augmented the baseline PLP features with MLP derived features coming from the combination of HATs and a conventional MLP using PLP as inputs. Both HATs and the conventional MLP are trained using the same training set of 32 hours through the same procedure in [2]. Each MLP has about 500K parameters. The MLP augmented feature achieves a WER of 37.5%, a 10% reduction from the baseline PLP feature.

The HTK-based system used decision-tree tied state triphone HMMs made by the following procedure: 1) train a monophone model, 2) convert the monophone model to a triphone model, 3) cluster the triphone states, 4) increase the number of Gaussian mixtures incrementally. Each time a modification was applied to the HMM such as the mixture splitting, the EM parameter estimation was iterated five times. We trained and evaluated the triphone HMMs with varying combination of the number of states and the number of Gaussians in each state. In our experiments using the tuning set, 2k and 1k states with 32 mixtures were the best for male and female models, respectively. The word error rates were 42.9% and 41.0% for the male and female test set, respectively, for the 1000-word vocabulary language model, and the average error rate on the full test set was 42.0%.

5. SELECTING DATA FOR TRAINING

If it is the case that some data are more useful for speech recognizer training than others, then more sophisticated data selection techniques may be of interest. Obviously, training

data that is mismatched to the testing condition is less useful than matched data; however, there is some evidence to suggest that even within matched collections there are more and less useful subsets of data. Early work in this area focused on using data identified as "useful" for refining an existing system [7]. In this work, a useful utterance was one which had a high word-error rate from an available recognition model, and low confidence (or likelihood) was used as an approximation for the case where the data was not already transcribed. Other work applied a similar algorithm to an utterance identification task in an active learning paradigm [8]. In this work, and in other work using a boosting approach to training [9], the idea is that utterances that are not accurately recognized (or well modeled) have a greater effect on improving classifier performance. This work differs from our problem in that their goal was to pick data to augment an initial set or improve an initial classifier, whereas our goal is to define the initial set for purposes of rapid (but meaningful) experimentation. Hence, we look at the question of whether it is useful to include more vs. less "outlier" type data in defining a small training set. Initial work at UW in small task definition suggested that *excluding* the outliers gave improved performance [10].

In the experiments below, we investigate this question with a variety of sampling techniques, with sampling based on average per-frame log-likelihood of forced alignments using the MFCC within-word triphone models of the Fall 2003 SRI system. (We can think of likelihood as a low-cost proxy for WER, but it may be justifiable in itself as an indicator of model fit.) In designing the sampling methods, we looked at the distribution of average log-likelihoods for the training data. As an example, the distribution for the male subset of the Fisher corpus is given below in Figure 1; distributions for females and for the Switchboard data were similar. We do not yet understand the reasons for the bimodal distribution, though we have observed that short utterances tend to have low log-likelihood. However, the fact that the data has a bimodal distribution impacts the strategy for avoiding or emphasizing outliers, which led us to consider the 2-mode sampling technique among others described below.

Subject to the constraints of obtaining balanced male/female sets and maintaining the same balance of Fisher and Switchboard data, six types of sampling were compared:

- *uniform:* every n-th utterance was selected;
- *trimmed:* based on the distribution of the average acoustic log-likelihoods, utterances within one standard deviation of the mean were selected, so outliers with either low or high log-likelihood were not used;
- *centered:* based on the distribution of the average acoustic log-likelihoods, utterances nearest to the mean were selected, so only utterances in a narrow range of typical log-likelihoods were used;
- *2-mode trimmed:* similar to *trimmed*, except that the two mode means are used;
- *2-mode centered:* similar to *centered*, except that the two mode means are used; and
- *flat:* utterances were selected so that the distribution of the average log-likelihood became flat in the subset, which effectively increases the emphasis on outlier utterances.

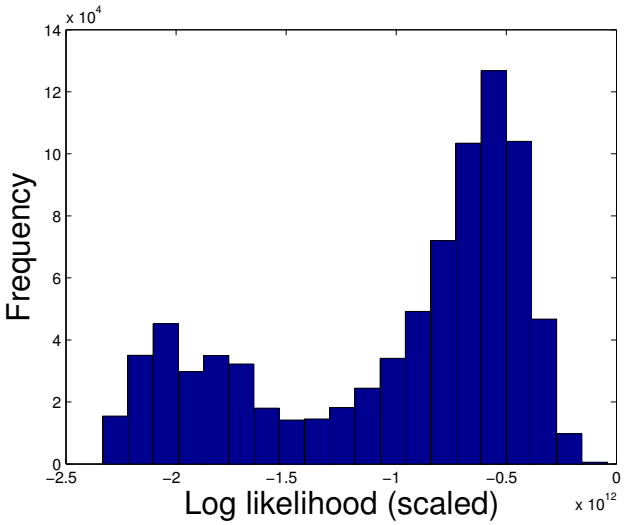

Fig. 1. Distribution of per-frame utterance log-likelihoods for the male subset of the Fisher corpus. Note the log likelihood is scaled for numerical reasons.

The different alternatives are illustrated in Figure 2. The different training subsets had the same amount of data and Fisher/Switchboard ratio as the baseline described in Section 3 and 4 in terms of duration and thus the number of utterances were not necessarily the same. In the 2-mode selection methods, the sampling was constrained so that the ratio of utterances in each mode was preserved. The selection by uniform sampling was exactly the same as the baseline, with the exception that in our case we started from the full Fisher and Switchboard sets (i.e. there was no restriction to native speakers). The various methods allowed us to explore whether the outlier utterances hurt or help in designing a small task.

The experiments were conducted using the 500-word vocabulary language model and the male test set, so error rates are slightly higher than those given earlier. The results are summarized in the middle column of Table 1. The *2-mode trimmed* sampling approach (avoiding outliers) gave the best results, and the flat sampling result had the worst

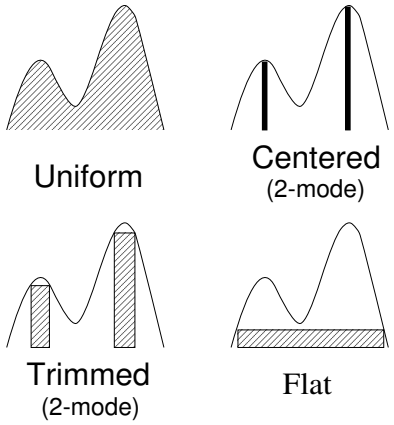

Fig. 2. Illustration of sampling alternatives explored.

Table 1. Recognition results for different training conditions.

Sampling	Stage 1	Stage 2
Uniform	45.5	45.0
Trimmed	45.4	–
Centered	45.6	–
2-mode trimmed	45.1	44.9
2-mode centered	45.5	45.0
Flat	45.9	45.3

performance. Other results were not significantly different from uniform sampling. These results suggest that avoiding rather than emphasizing outliers is the best data selection strategy for initial model design.

Given that we want the results to generalize to larger tasks, it seems important to include outliers in some way. Motivated by ideas from child language acquisition, which show that initial exposure usually involves more prototypical examples of speech sounds (measured by changes in formant frequencies) as mothers typically use when speaking to their infants [11], we decided to explore two-stage training experiments. The first stage would be based on more prototypical utterances, specifically the model generated from the best-case (2-mode trimmed) set of data. The variances in the model was then retrained with a larger data set, where the trimmed set is augmented by data using four possible sampling methods: uniform, trimmed, centered and flat. The idea was to use more prototypical data to obtain an initial model, which would then be improved with additional data from a broader distribution, hopefully resulting in a lower experimentation cost (and potentially better performance) than training with the full set to begin with. The different sampling conditions were again chosen to investigate the question of whether it is better to avoid or emphasize outliers. The results in the last column of Table 1 again show that the flat sampling actually hurts performance, and hence emphasizing outliers with a small training set is not a good idea.

Table 2. Recognition results for standard and 2-stage training.

Training	WER
Standard	45.3
2-stage (all)	45.1
2-stage (var)	44.9

The 2-stage training approach is faster than training with the combined data sets from the start, but we also wanted to consider whether there are performance advantages. Hence, we compared the standard training paradigm to 2-stage training and 2-stage training with only the variances adjusted in the 2nd stage (i.e. keeping the means or "prototypes" fixed). As can be seen in Table 2, we found that 2-stage methods give slightly better results than the standard training paradigm. For the 2-stage methods, the variance constrained training worked slightly better than the non-constrained training.

6. CONCLUSIONS

In this paper, we have described a small task for rapid experimentation on English CTS tasks and shown that results on such a task have generalized to a state-of-the-art system. Key to the task definition is the use of a small vocabulary of frequently observed words. The training and test definitions, as well as an HTK recipe for system design, will be available to other sites, with the hope that this will stimulate research in novel approaches to acoustic modeling.

In addition, we have investigated strategies for sampling a large corpus of data in order to best define a small training set. We found that excluding outliers, in the sense of very high or low forced alignment log-likelihood, is a useful strategy for designing the initial set. In addition, the 2-stage training method not only reduced the computation time but also gave slightly better performance compared to the standard training paradigm. However, it remains to be seen whether gains associated with the different training selection mechanisms will impact the transferability of experiment findings on the small task to a larger task and more complex systems. It may be simply the use of the frequent words (which are well-represented in a small training set) that is the key to transferability.

A potential problem with the sampling work is the use of likelihoods as a selection criterion, since low likelihoods can occur for several reasons. If low confidence is due to a

noisy signal or poor transcriptions, the utterances would not be good for training. Utterances with low confidence due to a dialect, for example, might be useful to include in training in order to represent a broader group of people. Further work with other criteria would be of interest.

Acknowledgments

The authors thank Andreas Stolcke for his advice in training the SRI recognition system for the small task. This work was supported by DARPA grant No. MDA972-02-1-0024. The opinions and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

7. REFERENCES

[1] N. Morgan, B. Y. Chen, Q. Zhu, and A. Stolcke, "TRAPping conversational speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition," in *Proc. Int'l Conf. on Acoustic, Speech and Signal Processing*, 2004, vol. 1, pp. 537–540.

[2] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP features in LVCSR," in *Proc. Int. Conf. on Spoken Language Processing*, 2004, vol. 2, pp. 921–924.

[3] S. Young *et al.*, *The HTK Book (for HTK Version 3.2)*, Cambridge University, 2002.

[4] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: an open source software system for speech and time-series ppocessing," in *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing*, 2002, pp. 3916–3919.

[5] T. Hain, "Implicit pronunciation modelling in asr," in *Proc. Pronunciation Modeling and Lexicon Adaptation Workshop*, 2002, pp. 129–134.

[6] A. Stolcke *et al.*, "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, 2000.

[7] T. M. Kamm and G. G. L. Meyer, "Selective sampling of training data for speech recognition," in *Proceedings of Human Language and Technology*, San Diego, CA, USA, 2002.

[8] S. Douglas, "Active learning for classifying phone sequences from unsupervised phonotactic models," in *Proceedings of HLT-NAACL*, Edmonton, Canada, 2003.

[9] G. Zweig and M. Padmanabhan, "Boosting gaussian mixtures in a LVCSR system," in *Proc. Int'l Conf. on Acoustic, Speech and Signal Processing*, Istanbul, Turkey, 2000, IEEE, pp. 1527–1530.

[10] O. Cetin, *Multi-rate Modeling, Model Inference and Estimation for Statistical Classifiers*, Ph.D. thesis, Univ. of Washington, 2004.

[11] P. K. Kuhl et al., "Cross-language analysis of phonetic units in language addressed to infants," *Science*, vol. 277, pp. 684–686, August 1997.