On the Interaction Between True Source, Training, and Testing Language Models*

Douglas B. Paul†, James K. Baker‡, and Janet M. Baker‡

†Lincoln Laboratory, MIT
Lexington, Ma. 02173

‡Dragon Systems, Inc.
90 Bridge St.
Newton, Ma. 02158

Abstract

An interaction has been found between the true source language model, training language model, and the testing language model. This interaction has implications for vocabulary independent modeling, testing methodologies, discriminative training, and the adequacy of our current databases for continuous speech recognition (CSR) development. The current DARPA databases suffer from the described difficulties which suggests that new CSR databases are needed if we are to further advance the state-of-the-art.

The Interaction During Training

When a category model (e.g. a context-free (CF) model such as a monophone) is used to a model a set of subcategories (e.g. context-dependent (CD) models such as triphones), the category model becomes the subcategory prior-probability weighted average of the subcategory models:

$$M_{cat} = \frac{\sum p_{subcat} M_{subcat}}{\sum p_{subcat}}$$

where M denotes a model. (The mathematics used here are intended to be conceptual rather than rigorous. Thus models will be considered to be averages. In practice, the method for deriving a model from a set of sub-models or observations is highly dependent upon the form of model used.) In a field, such as speech recognition, where models are trained from exemplars, the subcategory model will generally be:

$$M_{subcat} = \frac{1}{N} \sum_{i=1}^{N} O_{subcat,i}$$

where $O_{subcat,i}$ is an observation emitted from the subcategory. M_{cat} combines both the subcategory models and the prior-probability of the subcategories and similarly M_{subcat} combines the observations and their (sampled) prior-probabilities.

*This work was sponsored by the Defense Advanced Research Projects Agency.

In speech recognition, a phone category would contain some set of subcategories and a subcategory would defined by some specific set of context factors. There are many factors which may be used to define the subcategories [3]; a commonly used set is triphone [18] subcategories and monophone categories. Alternatively, stressed and unstressed phones might be combined. (Note that this averaging is recursive: subcategories are the combination of some set of subsubcategories and so on...)

We assume that speech is generated from some "true source" language model. (This language model would change as a function of many factors such as topic, history, and participants, but we will assume it to be constant for each task.) This true language model is known for some artificial tasks such as the DARPA Resource Management (RM) database [16], but can be estimated for naturally elicited speech and text if sufficient data is available. (However, current techniques for estimating language models are fairly rudimentary.)

Since the acoustic realization of the phones will be a function of this true language model, any acoustic models averaged over any group of subcategories will learn this true language model to some degree. (Learning the language model "to some degree" may be viewed as favoring the more likely subcategories.) Pragmatically, we have insufficient data to model all relevant subcategories separately and, even if we had sufficient data, we currently have insufficient computational resources to process all of it in any practical manner. Thus, since we must combine subcategories into larger models, a recognition system would favor the subcategories that were more commonly observed during training.

Implications for Performance Testing

Recognition is performed using some explicit language model. (No-grammar is a language model in which all following words are equally likely.) If the performance of a system is tested using a weaker language model than the true source language model, the acoustic models, if they have been affected by the training data language model, will strengthen the the total language model in the recognizer. Thus, one would expect better recogni-

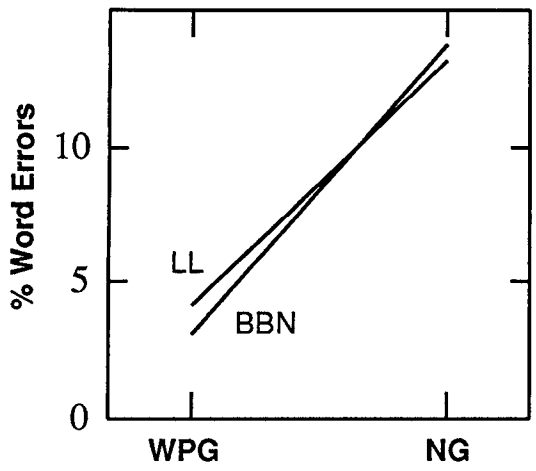

Figure 1: February 89 SD Evaluation Tests

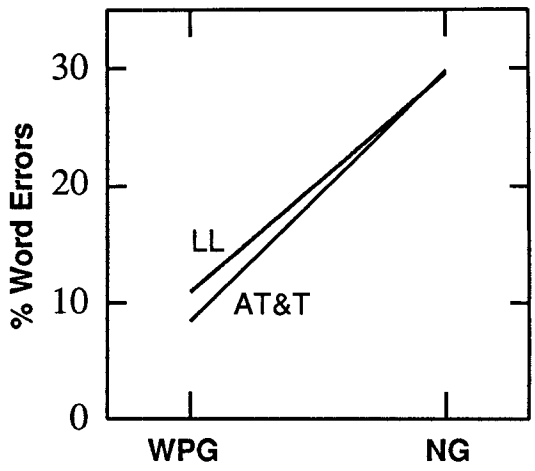

Figure 2: October 89 SI-109 Evaluation Tests

tion performance than would be predicted by the perplexity of the explicit recognition language model.

For example, the RM database was generated from a set of perplexity 9 patterns. The two official test conditions are a perplexity 60 word-pair grammar (WPG) and a perplexity 991 no-grammar (NG). But since the training data is perplexity 9, the acoustic training data includes a limited set of contexts. The explicit testing language model may allow a greater set of contexts, but the acoustic models in the recognizer are biased toward the contexts that are actually observed in the training data. WPG recognition performance would be expected to be worse if trained on data actually generated by the WPG. One would expect a similar effect if NG training data were used for NG tests. The net effect is that our performance testing is misleading: the performance obtained with the WPG does not tell us what the performance would be on a true perplexity 60 task. Similarly the NG tests do not tell us what the performance would be on a true perplexity 991 task.

In fact, rank orderings of systems tested using the RM database are not necessarily the same for WPG and NG tests. Figures 1 and 2 show the results of selected site-pairs from the February 89 speaker-dependent (SD) evaluation tests [11] and October 89 speaker-independent with 109 training speakers (SI-109) evaluation tests [6, 14]. The LL systems may be able to make better use of language model information stored in the acoustic models than are the other systems. (There are other possible explanations, but this one cannot currently be ruled out.)

Implications for Discriminative Training

Discriminative training (training for the right answer rather than for the "most accurate model") is performed in the context of the training data which consists of samples from true source language model. In some forms, including corrective training [2], it is performed in the additional context of an explicit training language model. (Corrective training uses a recognition pass to obtain possible confusions with the correct answer and the training language model is applied in this recognition pass. It is the only form of discriminative training that has been shown to improve performance on large vocabulary recognition tasks [2, 7].) Errors or near misses are used to perturb the acoustic models to lessen the possibility of error. But these errors are a strong function of the true source model of the training data and the training language model. Thus, these techniques increase the amount of the true source and training language models that are included in the acoustic models.

One of the stated advantages of discriminative training is that it corrects for an incorrect (form of) model. It does this by altering the trainable factors (parameters) of the model to account both for improper choice of function and for aspects of the true source which have not been included in the model. The second effect is exactly the above stated problem.

Evidence showing that corrective training inserts the training language model into the acoustic models has appeared in results reported using the CMU SPHINX system operating on the RM database. It was found that an NG corrective-trained set of models, which improved NG recognition performance, damaged WPG recognition performance compared to a maximum likelihood trained set of models [8].

Implications for Vocabulary Independence

The above suggests that any training methods that average over a number of contexts and/or use discriminative techniques include the true source language model in the acoustic models. Thus any set of acoustic models would be optimized for that specific task and therefore would be inferior when tested upon another task. CMU has investigated this problem and found RM models to yield poor performance on a different task [5]. We

have also performed some informal experiments which attempted to port RM triphones to our flight demonstration task (28 word vocabulary, perplexity 7 finite-state grammar) and found inferior performance compared to task-trained models.

What can be done?

Several things can be done to minimize the damage done to the acoustic models by the true source language model and any training language models. Since we have techniques which implement the language model independently of the acoustic model, it would help if the acoustic models were as free as possible from biases due to the training environment.

The first technique is to use a source of training data with a rich and realistic true language model. (The richness can be further increased by using data from a number of different sources.) This will provide a rich set of contexts to allow our systems to see the full range of contexts in which a real system will have to operate. CMU has already shown that a richer data source improves one's ability to produce task independent acoustic models [5].

A second technique is to minimize averaging across contexts in order to limit the ability of the acoustic models to model the language. A number of sites have already moved in this direction by changing from context-independent phone (monophone) models to left and right context dependent phone (triphone) models [18]. A further step along this line has been the inclusion of cross-word triphone models [7, 10, 13] which has minimized the ability of the acoustic models to learn the bigram language model.[1] These changes have improved recognition performance when trained and tested on the same database, but their effects on vocabulary independence have not been tested.

A third technique is to use larger training datasets. ("There is no data like more data."). This allows us to train more contexts and minimize the smoothing (averaging) required to train models. Comparisons of performance with increased amounts of training data using the RM1 database for speaker-independent training (109 vs. 72 speakers) and the RM2 database for speaker-dependent training (2400 vs. 600 sentences per speaker) have shown improved results [15] within task. The CMU vocabulary-independence experiments showed improved cross-task performance [5], but since the amount of data and the richness of the data were increased simultaneously, it is not currently possible to separate the two effects. We currently use less than ten hours of training data. In comparison, a typical school age child has heard thousands of hours of speech.

A fourth technique is to limit discriminative training to cases where the language model and vocabulary are known at training time. Discriminative training explicitly alters the acoustic models to reflect the true source and training language models. This may improve within task performance, but will not help and may impair cross-task performance. (The phone models have, in effect, been made fragile with respect to vocabulary and task.) However, it is not clear that this within-task advantage will be maintained with a good language model and better acoustic modeling techniques using non-discriminative training on adequate amounts and kinds of data.

A final technique is to test with a good language model. Testing with intentionally weakened language models asks the acoustic modeling to perform a task that it has not been trained to do. While the simplicity of no-grammar testing may be attractive, it is also the most misleading test condition and does not always predict the performance of a system with a good language model. If a realistic language model is used with properly trained acoustic models, the language model will perform the word sequence modeling and the acoustic models will perform the acoustic modeling without each trying to do the other's job.

Given an adequate amount of adequately rich data to train the acoustic models and an appropriate language model for the task, it should be possible to obtain good recognition performance using vocabulary and task independent acoustic models.

A Proposal for a Rich and Realistic DARPA CSR Database

None of the CSR databases currently available to the DARPA community meets all of the above requirements. The RM database was a good database for its time, but we have since found a number of weaknesses (and created some by improving our recognizers.) It was, however, a focal point around which much progress in CSR was achieved and if we design and produce a successor properly, we can initiate a similarly productive era. (The following proposed database serves a different purpose than and should be recorded in addition to the DARPA Air Traffic Information System (ATIS) database [4, 17].) A list of desirable features for a CSR database are:

1. Should be based on real human communication to insure richness and realism
   (a) Should be based on a large corpus of text or transcribed speech
   (b) Transcriptions should be available to the research community to allow language modeling.

[1] If sufficient training data is available to train the cross-word-context phone models without averaging, cross-word-context-dependent phone modeling will clearly minimize the effects of the bigram language model. However, the training data for the referenced systems was sufficiently limited that it was necessary to smooth (weighted average) some phone models for robustness and it was necessary to use cross-word-context-independent (i.e. averaged over observed word-boundary phones) phone models for the unobserved cross-word phone models needed by the recognizer. The net effect for limited training data is unknown—the cross-word phone models may increase the bigram language model learning for limited amounts of training data. Definitive experiments isolating this effect have not been reported.

2. A good standardized stochastic language model (or set of models)
3. A large amount of acoustic data
   (a) Both SD and SI training data
   (b) Standardized training, development test, and evaluation test data
   (c) A large number of test speakers to minimize the effects of speaker variation.
4. The task should be difficult enough to have a sufficiently high error rate even with a good language model.
   (a) We make the best progress when our tasks have a moderate error rate
   (b) We need enough errors for statistically significant experiments.
5. Extendibility to more difficult acoustic tasks to allow future growth (e.g. larger vocabularies).

The standardized language models and acoustic datasets are essential to allow rigorous inter-site system comparisons. If the language model training data is available, sites will also be able to work on improved language modeling.

The specific proposal is:

1. The ACL/DCI contains several large text databases [9]. The best one for our purposes is probably the transcriptions of Canadian parliamentary hearings part of the Canadian Hansard database. (Other viable alternatives are the parliamentary debate portion of the Hansard database or the Wall Street Journal texts.) This is the transcription of about 50M words of speech. It should be possible to derive a good bigram or trigram language model from this text and several other databases are available to facilitate cross-database language model investigations. The data is available to the research community.
2. 5000 words is probably a good vocabulary size given the current state of the art. We could use the 5000 most common words in the text database. If we set aside a block of about 10% of the text for CSR testing, we can obtain in-vocabulary CSR test sets and CSR test sets which include out-of-vocabulary words. This would also allow for extension to larger vocabularies when the CSR technology has improved sufficiently. The training sentences need not be limited to the chosen vocabulary.
3. The database would consist of read speech. This is fast and cheap to enable us record sufficient acoustic data. It would not attempt to cover extemporaneous speech phenomena. (Extemporaneous speech phenomena can be explored using the ATIS database. The ATIS database, however, does not have sufficient text backup to generate a good statistical language model.)

A cheaper, but less useful alternative is the continuous speech version of the IBM 5000 word vocabulary office correspondence (OC-5000) database [1]. This database has a good bigram language model, has sentence lists (but the test list is only 50 sentences), and has acoustic data which has already been recorded. However, the underlying text is not available and the availability of the language model and the acoustic data is currently uncertain due to unsettled legal issues [12].

Conclusion

Our training methods have been found to include a language model bias into our acoustic models. This bias causes misleading test results and impairs the vocabulary independence of our models. Richer and larger acoustic databases, context dependent modeling, avoiding discriminative training methods, and good testing language models all will serve to minimize this bias. A speech database based upon one of the ACL/DCI text databases would provide a good arena for continued CSR development while minimizing problems due to the training biases.

References

[1] A. Averbuch, L. Bahl, R. Bakis, P. Brown, A. Cole, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. de Souza, E. Epstein, D. Fraleigh, F. Jeninek, S. Katz, B. Lewis, R. Mercer, A. Nadas, D. Nahamoo, M. Picheny, G. Shichman, and P. Spinelli, "An IBM-PC Based Large-Vocabulary Isolated-Utterance Speech Recognizer," ICASSP 86, Tokyo, April 1986.

[2] L. R. Bahl, P.F. Brown, P. V. de Souza, and R. L. Mercer, "A New Algorithm for the Estimation of Hidden Markov Model Parameters," Proc. ICASSP 88, New York, NY, April 1988.

[3] F. R. Chen, "Identification of Contextual Factor For Pronunciation Networks," Proc. ICASSP90, Albuquerque, New Mexico, April 1990.

[4] C. Hemphill, "TI Implementation of Corpus Collection," this proceedings, June 1990.

[5] H. W. Hon and K. F. Lee, "On Vocabulary-Independent Speech Modeling," Proc. ICASSP90, Albuquerque, New Mexico, April 1990.

[6] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic Modeling of Subword Units for Large Vocabulary Speaker Independent Speech Recognition," Proceedings October 1989 DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, October 1989.

[7] K. F. Lee, H. W. Hon, and M. Y. Hwang, "Recent Progress in the SPHINX Speech Recognition System," Proceedings February 1989 DARPA Speech

and Natural Language Workshop, Morgan Kaufmann Publishers, February 1989.

[8] K. F. Lee and S. Mahajan, "Corrective and Reinforcement Learning for Speaker-Independent Continuous Speech Recognition," Technical Report CMU-CS-89-100, Carnegie-Mellon University, January 1989.

[9] M. Liberman, "Text on Tap: the ACL/DCI," Proceedings October 1989 DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, October 1989.

[10] H. Murveit, M. Cohen, P. Price, G. Baldwin, M. Weintraub, and J. Bernstein, "SRI's DECIPHER System," Proceedings February 1989 DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, February 1989.

[11] D. Pallett, "Speech Results on Resource Management Task," Proceedings February 1989 DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, February 1989.

[12] D. Pallett, personal communication, June 1990.

[13] D. B. Paul, "The Lincoln Continuous Speech Recognition System: Recent Developments and Results," Proceedings February 1989 DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, February 1989.

[14] D. B. Paul, "Tied Mixtures in the Lincoln Robust CSR," Proceedings October 1989 DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, October 1989.

[15] D. B. Paul, "The Lincoln Tied Mixture CSR," this proceedings, June 1990.

[16] P. Price, W. Fischer, J. Bernstein, and D. Pallett, "The DARPA 1000-word Resource Management Database for Continuous Speech Recognition," Proc. ICASSP 88, New York, April 1988.

[17] P. Price, The ATIS Common Task: Selection and Overview," this proceedings, June 1990.

[18] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," Proc. ICASSP 85, Tampa, FL, April 1985.