

Multi-source Training and Adaptation for Generic Speech Recognition*

Fabrice Lefevre, Jean-Luc Gauvain and Lori Lamel

Spoken Language Processing Group, LIMSI-CNRS, FRANCE
{lefevre,gauvain,lamel}@limsi.fr

ABSTRACT

In recent years there has been a considerable amount of work devoted to porting speech recognizers to new tasks. Recognition systems are usually tuned to a particular task and porting the system to a new task (or language) is both time-consuming and expensive. In this paper, issues in speech recognition portability are addressed and in particular the development of generic models for speech recognition. Multi-source training techniques aimed at enhancing the genericity of some wide domain models are investigated. We show that multi-source training and adaptation can reduce the performance gap between task-independent and task-dependent acoustic models, and for some tasks even out-perform task-dependent acoustic models.

1. INTRODUCTION

In the context of the EC IST-1999 CORETEX project we have investigated methods for developing systems with a high degree of genericity and adaptability. One main objective of this work is to develop *generic* speech recognition technology. By generic we mean a transcription engine that will work reasonably well on a wide range of speech tasks, ranging from digit recognition to large vocabulary conversational telephony speech, without the need for costly task-specific training data.

To start with the genericity of wide domain models under cross-task conditions (i.e., by recognizing task-specific data with a recognizer developed for a different task) has been assessed [6]. We chose to evaluate the performance of broadcast news acoustic and language models on three commonly used tasks: small vocabulary recognition (TI-digits), goal-oriented spoken dialog (ATIS), and read and spontaneous text dictation (WSJ). The broadcast news transcription task is rather general with a wide variety of linguistic and acoustic events encountered in the language. In addition, there are sufficient acoustic and linguistic training data available for this task so that accurate models covering a wide range of speaker and language characteristics can be estimated. These characteristics should lead to a reasonable coverage of the target tasks.

The objective of the work presented here is to obtain

generic models which are comparable or better than the respective task-dependent models for all tasks under consideration. To enhance the genericity of the models, we use training data from multiple sources to adapt the reference broadcast news models. Two main approaches are investigated for multi-source acoustic model estimation: pooled data adaptation and multi-step model adaptation. Complete model re-training is also contrasted with model adaptation. Multi-source language models are obtained by linear interpolation of the task-dependent models.

The next section provides an overview of the LIMSI broadcast news transcription system used as our reference system. In Section 3, the corpora used in this study are introduced. Multi-source training and adaptation techniques for acoustic modeling are proposed and tested in Section 4. Multi-source language model adaptation is described in Section 5 and experiments using multi-source acoustic and language models jointly are reported.

2. SYSTEM DESCRIPTION

The speech recognizer of LIMSI broadcast news transcription system [2] uses continuous density HMMs with Gaussian mixture for acoustic modeling and n -gram statistics estimated on large text corpora for language modeling. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained by means of a decision tree. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. The initial hypotheses are used for cluster-based acoustic model adaptation using the MLLR technique [8] prior to word graph generation. A 3-gram LM is used in the first two decoding steps. The final hypotheses are generated with a 4-gram LM and acoustic models adapted with the hypotheses of step 2.

In the baseline system used in DARPA evaluation tests, the acoustic models were trained on about 150 hours of audio data from the DARPA/LDC Hub4 Broadcast News corpus [5]. Gender-dependent acoustic models were built using MAP adaptation of SI seed models for wide-band and telephone-band speech [3]. The models contain 28000 position-dependent, cross-word triphone

*This work was partially financed by the European Commission under the IST-1999 Human Language Technologies project Coretex.

models with 11700 tied states and approximately 360k Gaussians [4]. The baseline 65k language models are obtained by interpolation of models trained on newspaper and newswire texts, commercial transcripts and transcriptions of acoustic training data.

The LIMSI 10xRT system had a word error of 17.1% on the 1999 NIST evaluation set and can transcribe unrestricted broadcast news data with a word error of about 20% [4].

3. CORPUS DESCRIPTION

For the small vocabulary recognition task, experiments are carried out on the adult speaker portion of the TI-digits corpus [9] (17k utterances from 225 speakers). The vocabulary contains the digits ‘1’ to ‘9’, plus ‘zero’ and ‘oh’. The database contains about 7 hours of high quality speech, equally divided between training and test. Our task-specific recognition system has only 108 context-dependent phone models due to the low phonemic coverage of the digits. The task-specific LM is a simple grammar allowing any sequence of up to 7 digits. In contrast to the BN system described above, word decoding is carried out in a single pass, and due to the short length of the utterances neither variance normalization of the cepstral coefficients nor speaker adaptation are performed.

The *DARPA Air Travel Information System* (ATIS) task is chosen as being representative of a goal-oriented human-machine dialog task, and the ARPA 1994 Spontaneous Speech Recognition ATIS-3 data [1] is used for testing purposes. The test data amounts to nearly 5 hours of speech from 24 speakers recorded with a close-talking microphone. Around 40h of speech data are available for training. The acoustic models used in our task-specific system include 1641 context-dependent phones with 4k independent HMM states. A trigram back-off language model was estimated on the transcriptions of the training utterances. The lexicon contains 1300 words, with compounds words for multi-word entities in the air-travel database (city and airport names, services etc.). Word decoding is carried out in a single trigram pass without speaker adaptation.

For the dictation task, the *Wall Street Journal* continuous speech recognition corpus [11] is used, abiding by the ARPA 1995 Hub3 test conditions. The acoustic training data consist of 100 hours of speech from a total of 355 speakers taken from the WSJ0 and WSJ1 corpora. The Hub3 baseline test data consist of studio quality read speech from 20 speakers with a total duration of 45 minutes. 21k context and position-dependent models were trained for the WSJ system, with 9k independent HMM states. A 65k-word vocabulary was selected and a trigram back-off model obtained by interpolating models trained on different data sets (training utterance transcriptions and newspapers texts). The word decoding procedure is the same as for the BN task.

For the reference BN transcription task, the conditions

| Task | New models | | Task-specific models |
|-------------------|-------------|------------|----------------------|
| | New config. | BN config. | |
| <i>BN (10×RT)</i> | 14.3 | 14.5 | 14.2 |
| <i>TI-digits</i> | 0.7 | 0.7 | 0.4 |
| <i>ATIS</i> | 4.1 | 3.1 | 4.1 |
| <i>read WSJ</i> | 7.3 | 7.2 | 7.6 |
| <i>spon WSJ</i> | 12.8 | 11.2 | 15.3* |

Table 1: Multi-source Training of Acoustic Models. Word error rates (%) for BN, TI-digits, ATIS and WSJ (read and spontaneous) test sets are given for three different configurations using task-specific lexica and LMs and MAP adapted BN acoustic models: (left) new models based a configuration obtained from the pooled data; (middle) new models using the BN configuration (i.e., CD phone list and state tying), and (right) task-specific acoustic models. (* Read WSJ models)

of the 1998 ARPA Hub4E evaluation [10] are followed. The system performance for the tasks of interest are reported in the last column of Table 1.

4. MULTI-SOURCE ACOUSTIC MODELS

In this section methods to improve acoustic model genericity via multi-source training and adaptation are investigated. The most straightforward approach consists of training new models using available data from all of the target tasks. Another approach is to adapt an existing model using data from the other tasks. Two adaptation schemes are investigated: the pooled data adaptation and the multiple step adaptation.

AM Training

Experiments with multi-source maximum likelihood training of acoustic models compared keeping the original BN configuration (i.e., the same context-dependent phone set and state-tying) with reselecting the phone contexts and reestimating the state-tying. The results are shown in the second and third columns of Table 1. Compared with the performance of the task-specific acoustic models, training new models on the pooled data yields essentially the same performance level when a new model configuration is derived, and globally improves the performance if the original BN configuration is kept. Better performances are observed for ATIS and WSJ (read and spontaneous) at the cost of a small degradation for BN. and TI-digits. The relative error reduction is 24% for ATIS, 5% for read WSJ and 27% for spontaneous WSJ.

Since significant improvements are observed with new models based on the BN model configuration, it is somewhat surprising that a new configuration specially derived for this experiment led to less good performance. An explanation for this result can be the special care with which the BN configuration has been obtained, resulting from a huge number of experiments carried out over the last years. In the experiments reported here, the new contextual model selection and state tying have been derived using the same thresholds as for the BN model training. Careful selection of the thresholds may lead to a better

balance between the amount of available training data and the number of independent states. As a consequence increasing the number of states (18401 in the new configuration vs. 11700 for BN) appears counterproductive. This is consistent with previous observations for BN where increasing the number of tied-states did not improve recognition performance.

AM Adaptation

Adaptation is investigated as an alternative to model re-training. The first adaptation procedure consists of pooling all of the available training data from all of the target tasks (the BN data is excluded). The pooled data is then used to adapt the BN acoustic models. As an alternative to data pooling, the BN models were sequentially adapted with data from the other tasks. Two task orderings were compared. In the first one, the BN acoustic models are first adapted with the WSJ data, then with ATIS data and finally with TI-digits data. In the second one, the order is reversed (TI, then ATIS, then WSJ).

The results for the different adaptation schemes are given in Table 2 using supervised MAP-based acoustic model adaptation [3]. All experiments are performed with task-specific lexica and language models. The BN system results reported in the table correspond to a system running in under 10xRT with the multi-source models used in both the 2nd and 3rd decoding steps¹. For the pooled data approach, the results are given for the global MAP adaptation of the models and for a contrastive setup where an additional MAP-adaptation to the specific task is added. Table 2 also reports the results for sequential adaptation using the two task orders.

Compared to the results obtained with task-dependent acoustic models, both the pooled data and the sequential adaptation schemes lead to better performance for ATIS and WSJ (read and spontaneous) at the cost of a small degradation for BN and TI-digits. For the pooled data approach, the introduction of a final adaptation for the specific task did not improve further the performance of the multi-source models (Table 2 ‘+task adapt.’). The adaptation results are comparable to those obtained with the training of new models. Read WSJ is the only task for which a further improvement is observed with adaptation.

The pooled data approach is seen to outperform sequential adaptation. Moreover, for the latter, the task order used for adaptation has a large influence on the resulting performance. With the order TI-digits→ATIS→WSJ, the performance with the sequential and pooling schemes are very close for ATIS and read WSJ. With the reversed order (WSJ used first) a substantial degradation is observed for both tasks (12% relative for ATIS and 10% relative for read WSJ). The TI-digits appear to be a special case: firstly it is the only target task for which multi-source training does not outperform task-specific train-

¹Previous results were reported using the multi-source acoustic models only in the final 4-gram decoding step of the BN system [7].

| Task | Pooled Data | | Sequential | |
|------------|--------------|------|------------|------|
| | +task adapt. | | (1) | (2) |
| BN (10×RT) | 14.9 | - | 15.8 | 15.3 |
| TI-digits | 0.7 | 0.5 | 0.6 | 1.3 |
| ATIS | 3.1 | 3.3 | 3.6 | 3.2 |
| read WSJ | 6.7 | 6.8 | 7.4 | 6.7 |
| spon WSJ | 11.8 | 11.7 | 12.4 | 11.5 |

Table 2: Multi-source Adaptation of Acoustic Models. Word error rates (%) for BN, TI-digits, ATIS and WSJ (read and spontaneous) test sets with four different configurations using task-specific lexica and LMs and MAP adapted BN acoustic models: (left) pooled data AM adaptation, (middle left) pooled adaptation followed by individual task adaptation and (middle right and right) sequential AM adaptation with two different sequences BN→WSJ→ATIS→TI (1) and BN→TI→ATIS→WSJ (2).

ing, and secondly, the word error on this task is dramatically affected by the task order in the sequential scheme (116% relative increase). When TI-digits is the last corpus used for adaptation, the error rates of the two multi-source training approaches are similar. When this corpus is the first one used for sequential adaptation, the performance approaches that obtained with the reference BN acoustic models (1.7%). In this case, the subsequent adaptations have annihilated the effect of the TI-digit adaptation.

These experiments show that multi-source adaptation can be successfully applied to improve the genericity of the reference acoustic models. Although comparable performance could be obtained with both of the proposed approaches, the sequential has the drawback of being sensitive to the chosen task adaptation order. To have more conclusive results, the multi-source models should be tested on data from other tasks having comparable characteristics as our target tasks. We have shown that putting together data from various tasks in an appropriate way can produce models which perform correctly on each of the tasks. To go further, we would like to show that these models would also perform well on data from unseen tasks as long as they share sufficient characteristics with one of the multi-source tasks. Insofar as data from such tasks are not available to us right now, some indications are given by the spontaneous WSJ results.

The better performance obtained for spontaneous WSJ using BN models instead of read WSJ models (see [7]) can be attributed to the modeling of the spontaneous nature of the journalist dictation which is better represented in BN than in read WSJ. Although the BN and read WSJ tasks share common characteristics, it is not clear which one is closest to spontaneous dictation: the read WSJ task because it is read newspaper texts or the BN task because it has more spontaneous speech. However, the multi-source acoustic models are seen to give the best performance so far on the spontaneous WSJ. Even more, if the multi-source acoustic models are used in combination with the BN LM rather than the WSJ LM, the word

| Task | Pooled Data | |
|------------|-------------|-------|
| | AM | AM+LM |
| BN (10×RT) | 14.9 | 17.5 |
| ATIS | 3.1 | 4.0 |
| read WSJ | 6.7 | 8.6 |
| spon WSJ | 11.8 | 11.2 |

Table 3: Multi-source Adaptation of Acoustic & Language Models. Word error rates (%) for BN, ATIS and WSJ (read and spontaneous) test sets for two different configurations using task-specific lexica and LMs and MAP adapted BN acoustic models: (left) pooled data acoustic model adaptation, (right) pooled data acoustic and language model adaptation.

error is reduced from 11.5% to 10.8%. Since no spontaneous WSJ data were included in the multi-source adaptation data, this result reflects the increased genericity of the multi-source adapted BN acoustic models.

5. MULTI-SOURCE LANGUAGE MODELS

The same types of approaches explored for multi-source acoustic modeling can be applied to multi-source language model development: training of new models or adaptation of a reference model. In this work we explore the estimation of multi-source LMs based on a linear interpolation of the task-specific LMs. No global normalization of the LM tokens has been performed although some non negligible differences exist². The union vocabulary size is 83k-words. The pronunciation variants across the task-specific vocabularies have been combined; the average number of variants is 1.2 per word. The interpolation weights were optimized via an EM estimation on a development text set containing an equal number of words from each task. The derived weights are 0.4 for BN, 0.2 for WSJ and 0.4 for ATIS for the trigrams and for the fourgrams (the ATIS trigram is used).

Recognition experiments were performed using the multi-source acoustic and language models jointly (see Table 3)³. The introduction of the multi-source LM is only favorable to the spontaneous WSJ. For all other tasks, an increase in the word error rate is observed compared to using the multi-source acoustic models alone. For BN and read WSJ the error rates are even higher than those of the task-specific models (+2.6% for BN and 1.9% for read WSJ). It seems that in the case of language models the merging of the task data leads to a greater confusability during recognition. The lack of global normalization could have an influence by having different n-gram corresponding to the same word sequence (due to the use of compound words in BN but not in WSJ).

²Three elements were accounted for in scoring: global BN mapping rules have been applied for all tasks, compound words have been split and periods associated with isolated letters (from BN vocabulary) were deleted for ATIS scoring.

³Due to its very simple language the TI-digits task has not been considered for this work.

6. CONCLUSIONS

In this paper, new insights have been gained on the genericity of state-of-the-art speech recognition systems. Four tasks spanning a range of complexities were considered with the objective of building a single set of acoustic and language models which would perform as well as the task-specific models on each task. Training of new models was contrasted with adaptation of reference models. In the latter case, models from the broadcast news task were chosen as reference models since they cover a wide range of acoustic and linguistic conditions.

Globally, multi-source acoustic model training and adaptation were shown to improve the model accuracy, yielding recognition performance comparable or better than that obtained with task-specific models. The results obtained by the various multi-source schemes explored in the paper are quite close, with no one approach outperforming the others on every task.

When multi-source acoustic and language models are used jointly, a loss in performance is observed for the BN and read WSJ tasks (compared with task-specific results). Although not totally responding to our initial objective (a set of generic models performing as well as the task-specific models), the multi-source adaptation approach appears a convenient way to design generic models.

REFERENCES

- [1] D. Dahl, M. Bates *et al.*, "Expanding the Scope of the ATIS Task : The ATIS-3 Corpus," *Proc. ARPA Spoken Language Systems Technology Workshop 1994*.
- [2] J.L. Gauvain, G. Adda, *et al.*, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *Proc. ARPA Speech Recognition Workshop 1997*.
- [3] J.L. Gauvain, C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains," *IEEE Trans. on SAP*, 2(2):291-298, April 1994.
- [4] J.L. Gauvain, L. Lamel, "Fast Decoding for Indexation of Broadcast Data," *Proc. ICSLP'00*.
- [5] D. Graff, "The 1996 Broadcast News Speech and Language-Model Corpus," *Proc. DARPA Speech Recognition Workshop 1999*.
- [6] F. Lefevre, J.L. Gauvain, L. Lamel. "Toward task-independent speech recognition," *Proc. IEEE ICASSP'01*.
- [7] F. Lefevre, J.L. Gauvain, L. Lamel. "Improving genericity for task-independent speech recognition," *Proc. ISCA Eurospeech'01*.
- [8] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, 9(2):171-185, 1995.
- [9] R.G. Leonard, "A Database for speaker-independent digit recognition," *Proc. ICASSP-84*.
- [10] D.S. Pallett, J.G. Fiscus *et al.* "1998 Broadcast News Benchmark Test Results," *Proc. DARPA Broadcast News Workshop 1999*.
- [11] D.B. Paul, J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *Proc. ICSLP'92*.