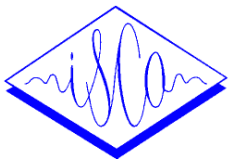

ITRW on Adaptation Methods
for Speech Recognitionn
Sophia Antipolis, France
August 29-30, 2001

Genericity and Adaptability Issues for Task-Independent Speech Recognition*

Fabrice Lefevre, Jean-Luc Gauvain and Lori Lamel

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

ABSTRACT

The last decade has witnessed major advances in core speech recognition technology, with today's systems able to recognize continuous speech from many speakers without the need for an explicit enrollment procedure. Despite these improvements, speech recognition is far from being a solved problem. Most recognition systems are tuned to a particular task and porting the system to another task or language is both time-consuming and expensive.

Our recent work addresses issues in speech recognizer portability, with the goal of developing generic core speech recognition technology. In this paper, we first assess the genericity of wide domain models by evaluating performance on several tasks. Then, transparent methods are used to adapt generic acoustic and language models to a specific task. Unsupervised acoustic models adaptation is contrasted with supervised adaptation, and a system-in-loop scheme for incremental unsupervised acoustic and linguistic models adaptation is investigated. Experiments on a spontaneous dialog task show that with the proposed scheme, a transparently adapted generic system can perform nearly as well (about a 1% absolute gap in word error rates) as a task-specific system trained on several tens of hours of manually transcribed data.

1. INTRODUCTION

In this paper we report on research carried out in the context of the EC IST-1999 CORETEX project, in which we are investigating methods for development of systems with high degree of genericity and adaptability. Genericity and adaptability refer to the capacity of the technology to work properly on a wide range of tasks and to dynamically keep models up to date using contemporary data. The more robust the initial generic system is, the less there is a need for adaptation. Unsupervised normalization and adaptation techniques should evidently be used to further enhance performance when the system is exposed to data of a particular type. One of our objectives is to develop *generic* core speech recognition technology. A generic transcription engine is one 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.

Adapting a speech recognizer to a new task or new language requires the availability of sufficient amount of transcribed training data. Since in the typical case acoustic data with detailed transcriptions are not available, the generation of such transcribed data is an expensive process in terms of manpower and time when changing recognition tasks. A proposed approach to reducing this investment is to use an existing recognizer (developed for other tasks or languages) to automatically transcribe the task-specific training data. These data can in turn be used to adapt the initial models to the new task or to directly improve the models genericity by means of multi-source training [5].

The genericity of wide domain models under cross-task conditions is assessed by using models developed for a task to recognize task-specific data from a different task. 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 (BN) task is quite general, covering a wide variety of linguistic and acoustic events. In the BN corpus distributed by the LDC there are sufficient acoustic and linguistic training data available that accurate models covering a wide range of speaker and language characteristics can be estimated.

The next section overviews the LIMSI broadcast news transcription system used as our generic system. In Section 3, cross-task experiments serve to gain insight about the degree of genericity of the BN models. Transparent adaptation techniques are shown to be effective in improving performance under cross-task conditions: Section 4 addresses acoustic model adaptation and Section 5 proposes a system-in-loop scheme for incremental unsupervised adaptation of both the acoustic and language models.

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,

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

3) final hypothesis generation. The initial hypotheses are used for cluster-based acoustic model adaptation using the MLLR technique [6] 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 BN system used in DARPA evaluation tests, the acoustic models were trained on about 150 hours of audio data from the DARPA Hub4 Broadcast News corpus distributed by the LDC. Gender-dependent acoustic models were built using MAP adaptation of speaker-independent seed models for wide-band and telephone band speech [3]. The models contain 28000 position-dependent, cross-word triphone models with 11700 tied states and approximately 360k Gaussians [2]. The baseline language models are obtained by interpolation of models trained on newspaper/newswire texts, commercial transcripts and transcriptions of acoustic training data. The recognition vocabulary contains 65120 words with an average of 1.2 pronunciations per word and represented with a set of 48 phones (including silence, filler words, and breath noises).

The LIMSI 10xRT system had a word error of 17.1% on the 1999 NIST evaluation set [9] and can transcribe unrestricted broadcast data with a word error of about 20% [2].

3. CROSS-TASK EXPERIMENTS

The development of a generic speech transcription engine first requires assessment of its performance across a range of tasks. This implies developing comparative task-specific systems, for which audio and textual data should be available for the targeted tasks. Three target tasks were selected among widely used corpora.

The TI-digits corpus [7] was selected for the small vocabulary recognition task. 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 content of the digits. The task-specific LM is a simple grammar allowing any sequence of up to 7 digits. Our task-dependent system has a WER of 0.4%, the best reported WERs on this task are around 0.2-0.3%.

The *DARPA Air Travel Information System* (ATIS) task [1] was chosen as being representative of a goal-oriented human-machine dialog task. Around 40 hours of speech data are available for training. The acoustic models used in our task-specific system contain 1641 context-dependent phones with 4k independent HMM states. A trigram back-off LM was estimated on the transcriptions of the 25k 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.). The word error rates for this task in the 1994 evaluation were mainly in the range of 2.5% to 5%, which we take as state-of-the-art for this task. The WER of our task-dependent system is 4.1%.

For the dictation task, the *Wall Street Journal* continuous speech recognition corpus [10] is used, abiding by the ARPA 1995 Hub3 test conditions. The acoustic training data consist of 100 hours of studio quality, read speech from a total of 355 speakers from the WSJ0 and WSJ1 corpora. The WSJ system has 21k context and position-dependent phone models, with 9k independent HMM states. The vocabulary contains 65k words and a trigram back-off model results from by interpolating models trained on different data sets (training utterance transcriptions and newspapers texts). The task-dependent system has a WER of 7.6% which is 1% higher than the best result reported at the time of the evaluation [12]. A contrastive experiment is carried out with the WSJ93 Spoke 9 data comprised of 200 spontaneous sentences spoken by journalists [4]. The WSJ system has a WER of 15.3% on these data. The best official result for this evaluation was 19.1% [11] but lower word error rates have since been reported on comparable test sets.

Task	*BN AMs & LMs*	*BN AMs & Task LMs*	*Task AMs & LMs*
BN	13.6	13.6	13.6
TI-digits	17.5	1.7	0.4
ATIS	20.8	4.7	4.1
WSJ read	11.6	9.0	7.6
WSJ spon	12.1	13.6	15.3

Table 1: Word error rates (%) for BN, TI-digits, ATIS, WSJ read and WSJ spontaneous test sets after recognition with three different configurations: (left) BN acoustic and language models; (center) BN acoustic models combined with task-specific lexica and LMs; and (right) task-dependent acoustic and language models.

For the reference transcription task, the conditions of the 1998 ARPA Hub4E evaluation [8] are applied. The 1998 evaluation, the LIMSI BN system had a WER of 13.6%.

Three sets of experiments are reported in Table 1. The first column shows the results of cross-task recognition experiments carried out using the BN acoustic and language models to decode the test data for the other tasks. The middle column gives results of experiments making use of mixed models, that is the BN acoustic models with task-specific LMs. The performances of the task-dependent models reported in the right column are close to the best reported results even though we did not devote too much effort in optimizing these models. It can also be observed by comparing the task-dependent (Table 1, right) and mixed (Table 1, middle) conditions, that the BN acoustic models are relatively generic. For TI-digits and ATIS the gap in performance is mainly due a linguistic mismatch since using using task-specific language models greatly reduces the error rate. For WSJ, the task-specific LMs are more closely matched to BN and only a 20% relative reduction in WER is obtained. On the spontaneous journalist dictation test data there is even an increase in WER using the WSJ LMs, which can be attributed to a better modelization of spontaneous speech effects (such as breath and filler words) in the BN models.

Task LMs	*BN Acoustic Models*		
Task	*Unadapted*	*Unsupervised*	*Supervised*
TI-digits	1.7	0.8	0.5
ATIS	4.7	4.5	3.2
WSJ read	9.0	6.9	6.5
WSJ spon	13.6	11.9	11.0

Table 2: Word error rates (%) for TI-digits, ATIS, WSJ read and WSJ spontaneous test sets after recognition with task-specific lexica and LMs and (left) BN acoustic models, (middle) unsupervised adaptation of the BN acoustic models and (right) supervised adaptation of the BN acoustic models.

Task	*BN models*	*Unsupervised Adaptation of BN Acoustic Models*
ATIS	20.8	13.5
WSJ read	11.6	7.8
WSJ spon	12.1	11.4

Table 3: Word error rates (%) for ATIS, WSJ read and WSJ spontaneous test sets with two recognition configurations using the BN lexicon and LMs: (left) BN acoustic models and (right) unsupervised global MLLR+MAP adaptation of the BN acoustic models.

4. ACOUSTIC MODEL ADAPTATION

The above experimental results show that while the reference BN acoustic models obtain relatively competitive results, the performances of task-specific models are better. Since one of our goals is to minimize the cost and effort in tuning to a target task, we are investigating methods to transparently adapt the BN reference acoustic models with task-specific data. By transparent we mean that the procedure is automatic and can be carried out without any human expertise. The approach proposed earlier is applied, that is, the reference BN system is used to transcribe the training data of the target task. This supposes of course that audio data have been collected. However, the data collection cost can be greatly reduced since no manual transcriptions are needed. The performance of the BN models under cross task conditions is well within the range for which the approximate transcriptions can be used for acoustic model adaptation.

The reference acoustic models are then adapted by means of a conventional adaptation technique such as MLLR and MAP. By adapting the reference models, there is no need to design a new set of models based on the training data characteristics. Adaptation is also preferred over the training of new models as it is likely that the new training data will have a narrower phonemic contextual coverage than the original reference models. The adaptation procedure is based on a combination of the MAP and MLLR techniques (more details can be found in [5]).

Cross-task unsupervised adaptation is evaluated for the three tasks. The entire WSJ training set (100 hours) and 15 hours of the 40 hours of the ATIS training data were transcribed using with the BN acoustic and language models. For TI-digits, the training data was transcribed using a mixed configuration, combining the BN acoustic models with the simple digit loop grammar since using a task-specific LM does not add any additional cost for this task.

In order to assess the quality of the automatic transcriptions, the system hypotheses were scored against the manually provided transcriptions of the training data. The resulting word error rates on the training data are 11.8% for WSJ, 27.8% for ATIS and 1.2% for TI-digits.

Table 2 reports the word error rates obtained with the task-adapted BN models for the four tasks. The recognition tests were carried out under mixed conditions, using the adapted BN acoustic models and task-dependent LMs. Unsupervised acoustic model adaptation is seen to improve performance in every case: TI-digits (53% relative), ATIS (4% relative), WSJ (23% relative) and spontaneous WSJ (11% relative). For completeness, the task-specific audio data and associated transcriptions were used to carry out supervised adaptation of the BN acoustic models. As expected, supervised model adaptation outperforms unsupervised adaptation for all tasks. The difference in performance is quite substantial for both the TI-digits (37% relative) and ATIS (29% relative) tasks. Smaller relative gains of about 5% are obtained for the spontaneous dictation task and for the read WSJ data. The gain appears to be correlated with the WER of the transcribed data: the difference is smallest for the WSJ task, where the difference in WERs using BN or task-sepcific models is quite a bit smaller than the performance differences observed for ATIS and TI-digits.

An improvement in performance due to acoustic model adaptation is also seen when BN language models are used, as shown in Table 3. For WSJ read and spontaneous, the WERs (7.8% and 11.4%, respectively) are even lower than those of the task-specific system (see Table 1). The linguistic proximity of the BN and WSJ tasks can largely explain these results. For the ATIS task, a relative WER reduction of 30% is observed with unsupervised acoustic model adaptation.

These results confirm our hypothesis that better performance can be obtained by adapting generic models with task-specific data instead of directly training task-specific models. The TI-digits task is the only task for which the best performance is still obtained using task-dependent models rather than BN models which have been adapted in a supervised manner. For the other tasks, the lowest WERs are obtained with the supervised adapted BN acoustic models: 3.2% for ATIS, 6.6% for WSJ and 11.0% for spontaneous WSJ.

5. SYSTEM-IN-LOOP ADAPTATION SCHEME

Based on the observations made from the above experiments, we investigate an incremental unsupervised adaptation scheme where a speech recognizer is used to annotate untranscribed data which is in turn used for adaptation. In this system-in-loop adaptation scheme, a first subset of the training data is automatically transcribed using the generic system. The acoustic and linguistic models of the generic

Adaptation Data		*Unsupervised Adaptation of*	
Amount	*WER*	*BN AMs*	*BN AMs & BN LMs*
0	-	20.8	20.8
15h	27.8	13.5	6.9
15h+26h	14.0	8.7	5.5

Table 4: Word error rates (%) for ATIS as a function of the amount of data used for unsupervised adaptation: (left) amount of and WER on the adaptation data; (right) WER on the ATIS test data adapting only the acoustic models (1st column) and both the acoustic and language models (2nd column). The 26 hours of data are transcribed using BN models adapted with the first 15h of data.

system are then adapted with these automatically annotated data and the resulting models are used to transcribe another portion of the training data. This procedure can be iterated as long as new data are available. One obvious use of this scheme is for online model adaptation in a dialog system.

This scheme is applied to the ATIS task. Acoustic model adaptation is based on a combination of MLLR and MAP. Language model adaptation is performed by means of a mixture model combining the BN 3-gram backoff model with a 3-gram backoff model estimated on the automatic transcriptions. The interpolation weight is determined by minimizing the perplexity of a set of development texts.

The results are presented in Table 4 as a function of the amount of adaptation data. About one-third (15 hours) of the ATIS training corpus was transcribed using the BN system, and the remaining 26 hours were transcribed using the adapted BN acoustic and language models. Model adaptation results in a 74% relative (15.3% absolute) reduction in word error rate. The first 15-hour subset accounts for 90% of this reduction even though it only represents one third of the final data set. The second iteration also gives a significant improvement with a relative WER reduction of 20%.

The performance improvement is shared between the acoustic and linguistic model adaptations. With only acoustic model adaptation both steps give a relative WER reduction of about 35%. Language model adaptation gives predominance to the LM estimated from the transcriptions since the best interpolation weight is 0.2 for the BN LM (for both steps). Given the low interpolation coefficient for the BN LM, we measured the contribution of this component by comparing performance to the LM estimated only on the automatic transcriptions. The interpolated language model has an 1% absolute gain (6.9% vs 7.8%) with the 15h adaptation set. The large improvement coming from LM adaptation is somewhat surprising, given the relatively high word error rate (almost 28%) of the generic system. It can be argued that this error rate may not be so damaging for acoustic model adaptation, since many of the phonemes are likely to be correctly recognized, and errors should be somewhat random. A similar explanation for the language model is plausible if errors occur mainly on unimportant words or are randomly distributed. At the current time we are analyzing the errors in an effort to increase our understanding.

6. CONCLUSIONS

This paper provides new insights on the genericity of state-of-the-art speech recognition systems, by testing a relatively wide-domain system on data from several tasks with different complexities. Models from the broadcast news domain were chosen for reference models as they cover a wide range of acoustic and linguistic conditions. These acoustic models are shown to be relatively task-independent as there is only a small increase in word error relative to the word error obtained with task-dependent acoustic models, when a task-dependent language model is used.

We demonstrate that unsupervised acoustic model adaptation can reduce the performance gap between task-independent and task-dependent acoustic models, and that supervised adaptation of generic models can lead to better performance than that achieved with task-specific models. Both supervised and unsupervised adaptation are less effective for the digits task, indicating that this is a special case.

Incremental unsupervised adaptation of both the broadcast news acoustic and language models is shown to be effective for spontaneous dialog transcription. A 75% relative error rate reduction is obtained on the ATIS task with a system-in-loop scheme. The word error rate of the task-adapted system using only the audio data is 5.5%, which can be compared with the 4.1% word error rate of the task-specific system trained on the same 40 hours of data using the manual transcriptions.

REFERENCES

[1] D. Dahl *et al.*, "Expanding the Scope of the ATIS Task : The ATIS-3 Corpus," *ARPA SLST Workshop*, 3-8, 1994.

[2] J.L. Gauvain, L. Lamel, "Fast Decoding for Indexation of Broadcast Data," *ICSLP'00*, **3**:794-798.

[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] F. Kubala *et al.*, "The Hub and Spoke Paradigm for CSR Evaluation," *ARPA SLST Workshop*, 9-14, 1994.

[5] F. Lefevre, J.L. Gauvain, L. Lamel, "Improving Genericity for Task-independent Speech Recognition," *EuroSpeech'01*.

[6] 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.

[7] R.G. Leonard, "A Database for speaker-independent digit recognition," *ICASSP'84*.

[8] D.S. Pallett, J.G. Fiscus *et al.* "1998 Broadcast News Benchmark Test Results," *DARPA Broadcast News Workshop*, 5-12, 1999.

[9] D. Pallett, J. Fiscus, M. Przybocki, "Broadcast News 1999 Test Results," *NIST/NSA Speech Transcription Workshop*, 2000.

[10] D.B. Paul, J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *ICSLP'92*.

[11] G. Zavaliagkos *et al.*, "Improved Search, Acoustic, and Language Modeling in the BBN BYBLOS Large Vocabulary CSR Systems," *ARPA SLST Workshop*, 81-88, 1994.

[12] P.C. Woodland, M. Gales *et al.*, "The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task," *ARPA Speech Recognition Workshop*, 1996.