Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 19 (2005) 345–363
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Genericity and portability for task-independent speech recognition

Fabrice Lefevre *, Jean-Luc Gauvain, Lori Lamel

Spoken Language Processing Group, LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

Received 22 December 2003; received in revised form 16 November 2004; accepted 24 November 2004
Available online 7 January 2005

Abstract

As core speech recognition technology improves, opening up a wider range of applications, genericity and portability are becoming important issues. Most of todays recognition systems are still tuned to a particular task and porting the system to a new task (or language) requires a substantial investment of time and money, as well as human expertise.

This paper addresses issues in speech recognizer portability and in the development of generic core speech recognition technology. First, the genericity of wide domain models is assessed by evaluating their performance on several tasks of varied complexity. Then, techniques aimed at enhancing the genericity of these wide domain models are investigated. Multi-source acoustic training is shown to reduce the performance gap between task-independent and task-dependent acoustic models, and for some tasks to outperform task-dependent acoustic models.

Transparent methods for porting generic models to a specific task are also explored. Transparent unsupervised acoustic model adaptation is contrasted with supervised adaptation, and incremental unsupervised adaptation of both the acoustic and linguistic models is investigated. Experimental results on a 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 rate) as a task-specific system trained on several tens of hours of manually transcribed data.

* Corresponding author. Tel.: +33 1 6985 8068; fax: +33 1 6985 8088.
E-mail addresses: lefevre@limsi.fr (F. Lefevre), gauvain@limsi.fr (J.-L. Gauvain), lamel@limsi.fr (L. Lamel).

0885-2308/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2004.11.001

1. Introduction

The last decade has witnessed significant advances in the capability and performance of speech recognizers. For instance todays state-of-the-art systems are able to transcribe unrestricted continuous speech from broadcast news data with acceptable performance for automatic indexation purposes (Garofolo et al., 2000). The advances can in part be attributed to the availability of large amounts of training data, which have allowed the estimation of more accurate and complex models. At the same time, the availability of faster and cheaper computational means have enabled the development and implementation of better training and decoding algorithms.

Despite the extent of recent progress, recognition accuracy is still extremely sensitive to the environmental conditions and speaking style: speaker characteristics, channel type and background noise have an important impact on the acoustic component of a speech recognizer, whereas the speaking style and the discourse domain have a large impact on the linguistic component. For instance, a system trained on a large vocabulary read–speech corpus such as the *Wall Street Journal* corpus (Paul and Baker, 1992) is unlikely to provide optimal performance on a dialog task such as Air Travel Information Services (ATIS) (Price, 1990). The commonly adopted approach for achieving state-of-the-art performance is to develop a system for each specific task which also usually entails acquiring the necessary linguistic resources. Given the very large number of potentially different situations in which speech recognizers can be used, lack of genericity is a flaw that hinders the widespread use of speech recognition technology.

In the context of the CORETEX project[1] we started investigating methods for fast system development, as well as development of systems with a high degree of adaptability and genericity. By *fast system development* we refer to both task portability and language support, where *task portability* refers to the capability of adapting the technology to a new task by exploiting limited amounts of domain-specific knowledge and *language support* concerns the capability of porting technology to different languages at a reasonable cost. *Adaptability* is the capacity of the technology to dynamically keep models up-to-date using available data, and *genericity* refers to the capacity of the technology to work properly on a wide range of tasks without specific training prior to use. Concerning the acoustic modeling component, genericity implies that recognition performance does not vary too much as a function of the acoustic characteristics of the data (microphone, bandwidth, acoustic environment), or as a function of the speaker and the speaking style. Transparent normalization and adaptation techniques evidently can be used to further enhance performance when the system is exposed to data of a particular type.

This paper first assesses the genericity of wide domain models under cross-task conditions, i.e., by processing task-specific data with a recognizer developed for a different task. A broadcast news (BN) transcription system was chosen for the generic reference system. The BN task is relatively wide domain in that it covers a range of different acoustic and linguistic conditions: speaking styles ranging from planned to spontaneous speech; a wide variety of native and non-native speak-

[1] EC IST-1999 Coretex Project http://coretex.itc.it.

ers with different accents; close-talking microphones and telephone channels; recording conditions in quiet studios and on-site reports in noisy places; musical or other backgrounds; and a wide variety of topics. The performance of broadcast news acoustic and language models is evaluated on three "target" tasks covering a range of complexities: small vocabulary recognition, prepared and spontaneous text dictation, and goal-oriented spoken dialog. After demonstrating the inherent level of genericity of these models, different ways of using annotated task-specific training data from multiple sources to further enhance the genericity of these models is explored. The objective is to develop a single set of generic acoustic models which perform as well or even better than the respective task-dependent models for all tasks under consideration. To do so, two approaches for acoustic model adaptation are evaluated: pooling the data from the target tasks and multi-step sequential 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.

Next, porting the BN system to the three target tasks is investigated. The commonly used approach to develop a system for a new task (or another language) requires the availability of sufficient amounts of transcribed acoustic training data, as well as substantial amounts of task-related texts for linguistic model estimation. When changing to a new domain, detailed transcriptions of acoustic data are usually unavailable and must be produced. However, transcribing the data is an expensive process in terms of manpower and time. Recent studies (Kemp and Waibel, 1999; Lamel et al., 2000; Zavaliagkos and Colthurst, 1998) have investigated ways to reduce the development cost by using information with different levels of completeness to provide supervision. The same basic idea is applied here using the available task-specific training data in an unsupervised manner. Cross-task unsupervised adaptation is carried out using unannotated audio training data for the target tasks. Since no manual transcription is required, this approach is much less costly than standard task-specific training in terms of human effort. If carried out in an incremental manner, the speech corpus for the new domain can be cumulatively extended over time without direct manual transcription.

The remainder of this paper is organized as follows. The next section provides an overview of the transcription system and presents the corpora and task-specific systems used in this work. In Section 3, the genericity of the BN system is assessed via cross-task recognition experiments and genericity improvements are sought by means of multi-source training and adaptation. Portability of the reference BN system is investigated in Section 4 where the impact of acoustic and linguistic model adaptation are studied separately.

2. System and task descriptions

In this section, the transcription system used in this work and the task-specific systems developed using audio and textual data available for each target task are described. The reference system is a system developed for the broadcast news transcription task using the training materials distributed by the Linguistic Data Consortium (LDC) for use in the DARPA Hub4 evaluations. Three target tasks, spanning a range of complexities for which widely used corpora are available, were selected: small vocabulary recognition (TI-digits); dictation (WSJ); goal-oriented human–machine spoken

Table 1
Brief description and reference word error rate (%) for the task-specific corpora used in this work

Task	Corpus	Training material		Test material		
		Audio (#spk)	Textual	Eval set	Audio (#spk)	Ref. WER
TV and radio news transcription	Broadcast news Hub4 English	150 h (112)	Closed-captions, commercial transcripts, transcriptions of audio data	BN 98	3 h (113) (Pallett et al., 1999)	13.5
Small vocabulary	TI-digits	3.5 h (112)	–	Adult Portion	4 h (113) (Normandin et al., 1994)	0.2
Human–machine spontaneous dialog	ATIS	40 h (674)	Transcriptions of audio data	SPREC 94	1 h 30 min (24) (Dahl et al., 1994)	2.5
News dictation	WSJ	100 h (355)	Newspaper, newswire, transcriptions of audio data	CSR 95	45 min (20) (Woodland et al., 1995)	6.6
Spontaneous journalist dictation	*same as News dictation task*			CSR S9 93	43 min (10) (Zavaliagkos et al., 1994)	19.1

dialog (ATIS). The tasks and corpora characteristics are summarized in Table 1 along with a description of the evaluation data used in this work.

The speech recognizer uses continuous density hidden Markov models (HMMs) with Gaussian mixture for acoustic 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. The acoustic analysis derives cepstral parameters from a Mel frequency spectrum estimated on the 0–8 kHz band every 10 ms. Cepstral mean removal and variance normalization are applied to the cepstral coefficients. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with their first and second derivatives. Gender-dependent acoustic models are estimated using MAP adaptation of speaker-independent seed models for wideband and telephone band speech (Gauvain et al., 1994).

N-gram probabilities estimated on the task-related texts are used for language modeling, with each recognition vocabulary comprised of the most frequent words in the training texts. The TI-digit task is an exception since the recognition vocabulary is defined a priori and the language model consists of a simple digit loop grammar. A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. The pronunciations make use of a set of 48 phones, where 3 of the phone units represent silence, filler words, and breath noises. The filler and breath phones model only these events and are not used in the pronunciations of the other lexical entries. These two phones are not modelled in the TI-digit and WSJ tasks.

The broadcast news transcription task was retained for the reference task, with the BN system serving as the ''reference'' system. The BN transcription system has two main components, an

audio partitioner[2] and a speech recognizer. In this work the partitioning procedure is only used for the BN task.[3]

For the BN transcription system a multi-pass decoding strategy is used to optimize the performance computation time tradeoff (Gauvain et al., 2000). Recognition is performed in three decoding passes (for a total real-time factor under 10× real-time): (1) initial hypothesis generation, (2) word graph generation, (3) final hypothesis generation. After each decoding pass the word hypotheses are used for cluster-based acoustic model adaptation using the MLLR technique (Leggetter and Woodland, 1995). Different sets of acoustic models are used in each decoding step. A small set of acoustic models (5k triphone contexts, 6k tied-states) with 16 Gaussians per state is used in the first decoding step to generate the initial hypotheses. Then a larger set of models (28k triphone contexts, 11k tied-states) with 16 Gaussians per state is used to generate the word graph. The final decoding step uses a 4-g LM with a 32 Gaussians per state version of the larger model set.

For testing purposes, the experimental conditions of the 1998 ARPA Hub4E evaluation (Pallett et al., 1999) are adopted. The acoustic models are trained on about 150 h of audio data from the Hub4 Broadcast News corpus (the LDC 1996 and 1997 Broadcast News Speech collections) (Graff et al., 1997). The BN language models are obtained by interpolation of models trained on various text sources (excluding the test epochs): about 790M words of newspaper and newswire texts; 240M word of commercial broadcast news transcripts; and the transcriptions of the Hub4 acoustic data. The pronunciation dictionary contains 65,120 words with 76,644 phone transcriptions, and has a lexical coverage of over 99% on all evaluation test sets from the years 1996–1999. The lexicon includes about 300 compound words for frequent word sequences, as well as word entries for 1000 common acronyms, thus providing a way to allow for reduced pronunciations (Gauvain et al., 1997). In the 1998 evaluation, the LIMSI BN system had a word error of 13.6%.

The widely used TI-digits corpus (Leonar, 1984) was selected for the small vocabulary recognition task. This corpus contains about 7 h of high quality speech, equally divided between training and test. Due to the restricted phonemic content of the digits, the task-specific recognition system has only 108 contexts-dependent tied-state phone models for each gender. The task-specific LM is a simple grammar allowing any sequence of up to 7 digits. In contrast to the reference 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 task-dependent system built for this task has a word error of 0.4%. The best reported word error rates on this task are around 0.2–0.3% (see, for instance Gauvain and Lee, 1992b; Normandin et al., 1994).

The DARPA Air Travel Information System (ATIS) task is representative of a goal-oriented spontaneous dialog task and has been used for comparative evaluations (Dahl et al., 1994; Price, 1990). About 40 h of speech data are available for training. The acoustic models in the task-specific system contain 1641 context and gender-dependent tied-state phone models with 4k

[2] Data partitioning (Gauvain et al., 1997) serves to divide the continuous audio stream into homogeneous segments, associating appropriate labels with the segments. The result of the partitioning process is a set of speech segments with speaker cluster, gender and telephone/wideband labels.

[3] This partitioning step is required when dealing with "real world" data in the form of a continuous audio stream. In contrast, "laboratory" data are generally already segmented into individual utterances.

independent HMM states. A trigram back-off LM was estimated on the transcriptions of the 25k training utterances. The lexicon contains 1300 words with 1773 phone transcriptions, with some 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. The reported 1994 evaluation word error rates were mainly in the range of 2.5–5%, which are taken as reflecting state-of-the-art performance for this task. The word error rate of our task-dependent system is 4.1%.

For the dictation task, the Wall Street Journal (WSJ) continuous speech recognition corpus (Paul and Baker, 1992) is used, following the ARPA 1995 test conditions. The acoustic training data consist of about 100 h of studio quality, read speech from 355 speakers (WSJ0 and WSJ1 corpora). The task-specific WSJ system has 21k context, position and gender-dependent phone models, with 9k independent HMM states. The recognition vocabulary contains 65k words with 77k phone transcriptions. The trigram back-off language model is the result of interpolating models trained on different data sets (acoustic data transcriptions and newspapers texts). The word decoding procedure is the same as for the reference BN task. The task-dependent system has a word error of 7.6% which is about 1% higher than the best result reported at the time of the evaluation (Woodland et al., 1995). A contrastive experiment for the dictation task is carried out on the 1993 CSR Spoke 9 test data comprised of 200 spontaneous sentences spoken by journalists who dictated news articles on selected topics (Kubala et al., 1994). The WSJ system has a word error rate of 15.3% on this data. The best official result reported in the 1993 evaluation was 19.1% (Zavaliagkos et al., 1994). However, lower word error rates have since been reported on comparable although different test sets.

3. Assessing and improving genericity

This section reports on contrastive experiments carried out to assess the performance of the reference BN system when confronted with data from several tasks, and investigates multi-source training techniques as a way to increase the level of genericity of the acoustic models.

3.1. Cross-task recognition

Three sets of experiments are reported. The first is a set of cross-task recognition experiments using the reference BN system to decode test data for the three target tasks. To assess the respective contributions of the acoustic and language models, the second set of experiments make use of a mixed configuration, that is the BN acoustic models combined with the task-specific LMs. The third set of experiments using task-specific models serves as a comparative performance baseline. For the three experiments, recall that the spontaneous dictation task has no specific training data and so the task-specific models are those of read WSJ.

The word error rates obtained for the three sets of recognition experiments are reported in Table 2. A comparison with the rightmost column of Table 1 shows that the performance of the task-specific models are close to the best reported results even though not too much effort was devoted to optimizing these models. By comparing the task-specific (Table 2, right) and mixed (Table 2, middle) conditions, it can be seen that the BN acoustic models are somewhat generic.

Table 2
Cross-task recognition

Task	Cross-task	Mixed	Task-specific
	BN acoustic and language models	BN acoustic models and task language model	Task acoustic and language models
BN ($10 \times$ RT)	14.2	14.2	14.2
TI-digits	17.5	1.7	0.4
ATIS	17.8	4.7	4.1
Read WSJ	11.6	9.0	7.6
Spontaneous dictation	12.1	13.6[a]	15.3[a]

Word error rates (%) for BN, TI-digits, ATIS, read and spontaneous WSJ test sets are given for three different recognition configurations: (cross-task, left) BN acoustic and language models; (mixed, center) BN acoustic models combined with task-specific lexica and LMs; and (task-specific, right) task acoustic and language models.
[a] Task-specific models = Read WSJ models.

This is particularly true for the ATIS and read WSJ tasks where the performance loss with the BN acoustic models are 0.6% and 1.4%, respectively (12% and 18% relative) compared to the performance with task-specific acoustic models. The BN acoustic models seem to be a good starting point for building generic acoustic models.

For the TI-digits and ATIS tasks, the mixed condition results show that the gap in performance is mainly due to a linguistic mismatch. For both tasks, more than 90% of the degradation observed in the cross-task conditions can be imputed to the language models. The read WSJ and BN language models are more similar and the degradation due to the LM amounts to 65% of the total loss between cross-task and task-specific conditions. On the journalist dictation test data, there is even a reduction in word error using the BN LMs instead of the WSJ LMs, which can be attributed to a better modelization of spontaneous speech effects (such as breath and filler words) by the BN models.

These observations are supported by measures of the test set perplexity and out-of-vocabulary (OOV) rate for all the tasks. Table 3 reports the perplexities and OOV rates measured on the evaluation sets with the BN and task-specific language models. While TI-digits and ATIS have low OOV rates with both LMs, a large increase in perplexity is observed with the BN language model compared to the task-specific LMs. The perplexity increases from 11 to 137 for TI-digits and from 16 to 311 for ATIS. In contrast, for the read WSJ and the spontaneous dictation test sets, while the OOV rate is more than doubled between the BN and WSJ lexica, the perplexity of the trigram language model is higher for the read data (180 vs. 138) but is comparable for spontaneous dictation.

3.2. Multi-source acoustic models

In this section, methods to improve acoustic model genericity via multi-source training are investigated. The objective is to obtain generic models which give comparable or better performance than the respective task-dependent models for all tasks under consideration. The most straightforward approach consists of training new models using available data from all of the

Table 3
Cross-task recognition

Task	*N*-gram	BN LM		Task-specific LM	
		PP	OOV	PP	OOV
BN	3 g	164	0.3	164	0.3
	4 g	148	0.3	148	0.3
TI-digits	3 g	137	0.0	11	0.0
ATIS	3 g	311	0.1	16	0.1
Read WSJ	3 g	180	1.8	138	0.8
	4 g	149	1.8	124	0.8
Spontaneous dictation[a]	3 g	183	2.3	182	1.0
	4 g	167	2.3	170	1.0

Perplexities and OOV rates (%) of the evaluation sets are given with two different language models and associated lexica: (left) BN LMs and (right) task-specific LMs. Perplexities are given for trigram models (3 g) and fourgram models (4 g) when used.
[a] Task-specific models = Read WSJ models.

target tasks. Another approach is to adapt an existing model using data from the other tasks. Two adaptation schemes are investigated: pooled data adaptation and multiple step adaptation.

3.2.1. Multi-source acoustic model training

Experiments are reported with multi-source maximum likelihood training of acoustic models using all of the available data from the four data sources. The left part of Table 4 compares recognition performance keeping the original BN model structure (i.e., the same context-dependent phone set and state-tying) with reselecting the phone contexts and reestimating the state-tying on the pooled data. Training acoustic models with the revised model structure obtained with the

Table 4
Multi-source acoustic model training (task-specific LMs)

	Model training			BN model adaptation			
	Task specific	Multi-source		Pooled data		Sequential	
		New structure	BN structure	Global + task adaptation		Order 1	Order 2
BN	14.2	14.3	14.5	14.91	–	15.8	15.3
TI-digits	0.4	0.7	0.7	0.7	0.5	0.6	1.3
ATIS	4.1	4.1	3.1	3.1	3.3	3.6	3.2
Read WSJ	7.6	7.3	7.2	6.7	6.8	7.4	6.7
Spontaneous dictation	15.3*	12.8	11.2	11.8	11.7	12.4	11.5

Word error rates (%) for the 5 tasks under consideration using a common set of multi-source acoustic models which is either ML trained or adapted from the BN seed models. Three conditions are reported in this table, all using task-specific lexica and LMs: multi-source training with 2 model structures (columns 2 and 3), global model adaptation (alone in column 4 and followed by a task-specific adaptation in column 5) and sequential model adaptation following two opposite orders (order 1: WSJ, ATIS, TI-digits, and order 2: TI-digits, ATIS, WSJ). These results are contrasted with the error rates obtained using task specific models given in column 1.

pooled data (second column) yields essentially the same performance levels as the task-specific acoustic models (first column). Keeping the original BN model structure and training models on the pooled globally improves performance (third column). Better performances are observed for the ATIS, read WSJ and spontaneous dictation tasks at the cost of a small absolute degradation for BN and TI-digits (but a large relative degradation fot this latter). The relative error reduction is 24% for ATIS, 5% for WSJ and 27% for spontaneous dictation.

Since significant improvements are observed with multi-source training based on the BN model structure, it is somewhat surprising that the revised structure specially derived for this experiment led to less good performance. An explanation for this result can be the special care with which the original BN configuration was obtained, resulting from a large number of experiments carried out over the last few years. In the experiments reported here, the revised contextual model selection and state tying were derived using the same thresholds as for the BN model training. With approximatively double the amount of training data, the number of contextual phone models increased from 28,064 to 32,550 and the number of tied-states from 11,700 for BN to 18,401 in the revised structure. As a consequence increasing the number of states appears counterproductive. This is consistent with our previous experience with BN where increasing the number of tied-states did not improve recognition performance. Careful selection of the thresholds may lead to a better balance between the amount of available training data and the number of independent states.

3.2.2. Multi-source acoustic model adaptation

Adaptation is now investigated as an alternative to training acoustic models from scratch. The first adaptation procedure consists of pooling the available data from the three target tasks, and using the pooled data to adapt the BN acoustic models. As an alternative to data pooling, the BN models can be sequentially adapted with data from the target tasks. While the first approach can appear quite natural, the latter is proposed as a means to compensate for large variations in the training corpus size. Adapting separately for each task allows them to have equal importance irrespective of their amount of training material. In this case, however, the task order during adaptation process becomes an issue.

The results for the different adaptation schemes are given in the right part of Table 4 using supervised MAP-based acoustic model adaptation of the reference BN models (Gauvain et al., 1994). All experiments are performed with both task-specific lexica and language models. The multi-source models are used in the 2nd and 3rd decoding steps.[4] For the pooled data adaptation, results are given using global MAP adaptation of the reference models and for a contrastive setup where an additional MAP-adaptation to the specific task is added. In both cases, the model structure remains the same. No decision tree adaptation, as suggested in (Schultz and Waibel, 2001), has been performed. Table 4 also reports the results for sequential adaption using two task orderings. In the first one, the reference acoustic models are first adapted with WSJ data, then with ATIS data and finally with TI-digits data. In the second one, the order is reversed (first TI, then ATIS, then WSJ).

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, WSJ and spontaneous

[4] Previous reported results used the multi-source acoustic models only in the final 4-g decoding step of the BN system (Lefevre et al., 2001).

dictation at the cost of a small degradation for BN and TI-digits (except for order 2 in sequential adaptation). It is interesting to see that for the pooled data approach, the introduction of a final task-specific adaptation does not improve upon the performance of the multi-source models (Table 4, column 5). The results with adaptation are comparable to those obtained by training of new models. WSJ is the only task for which a larger improvement is observed with model adaptation.

Adaptation using the pooled data is seen to outperform sequential adaptation. Moreover, the task order used for sequential adaptation has a large influence on the resulting performance. With the second order (TI-digits, ATIS, WSJ), the performance with the sequential and pooling schemes are very close for ATIS and 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 task appears to be a special case: firstly it is the only target task for which multi-source training does not outperform task-specific training, and secondly, the word error on this task is dramatically affected by the task order in the sequential scheme (116% relative increase).

These experiments demonstrate that multi-source adaptation can be successfully applied to improve the genericity of the reference acoustic models. Although comparable performance can be obtained with both of the proposed adaptation schemes, the sequential one has the drawback of being sensitive to the task adaptation order. To have more conclusive results, the multi-source models should be tested on data from other tasks not used in training them.

The journalist dictation task, however, provides some indications on model genericity.

The better performance obtained for spontaneous dictation using BN models instead of read WSJ models (see Lefevre et al., 2001) can be attributed to better modeling of the spontaneous nature of journalist dictation. Although the BN and WSJ tasks share common characteristics, it is not clear which one is closest to spontaneous dictation. The multi-source acoustic models are seen to give the best performance on the spontaneous dictation. In addition, if the multi-source acoustic models are used in combination with the BN LM (rather than the WSJ LM), the word error rate is reduced from 11.2% to 10.8%. Since no spontaneous dictation data are included in the multi-source adaptation data, these results reflect the increased genericity of the adapted BN acoustic models.

3.3. Multi-source language models

In this section, multi-source language modeling based on a linear interpolation of the task-specific language models is investigated. Task-specific language models estimated on the BN, ATIS, and WSJ training texts are interpolated at the *n*-gram level.[5] The combined vocabulary, which is the union of the vocabularies for the three tasks, contains 83k-words. It should be noted that the task specific language models were optimized independently, no global normalization of the LM tokens has been performed even though some non-negligible differences exist, concerning in particular the use of compound words and acronyms.[6] The pronunciation lexicon includes all variants found in the task-specific lexica and has an average of 1.2 variants per word. The interpolation weights for the multi-source LMs were optimized via an EM estimation so as to minimize

[5] Due to its very simple language the TI-digits task has not been considered for this part of the study.

[6] Three differences 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.

Table 5
Multi-source language model adaptation

Task	Task specific LM	Multi-source LM
BN ($10 \times$ RT)	14.9	17.5
ATIS	3.1	4.0
Read WSJ	6.7	8.6
Spontaneous dictation	11.8	11.2

Word error rates (%) on the BN, ATIS, read WSJ and spontaneous dictation test sets for two different configurations using BN acoustic models adapted with the pooled data with task-specific LMs (left) and the interpolated multi-source LM (right).

the perplexity of a development text set containing an equal number of words from each task. The estimated trigram and fourgram weights are 0.4 for BN and ATIS and 0.2 for WSJ (for ATIS only a trigram LM is used).

Recognition results using the multi-source acoustic models with the multi-source language models are given in Table 5. The multi-source LMs are seen to reduce the word error only on the spontaneous dictation task. For all other tasks, an increase in the word error rate is observed. For BN and read WSJ the error rates are higher than those obtained with the task-specific models (+17% relative for BN and +28% for WSJ). It seems that interpolating language models from multiple tasks leads to greater confusability during recognition. The lack of global normalization for compound words and acronyms may cause errors in that different *n*-grams can correspond to the same word sequence. The straightforward use of a common normalization would, however, entail optimizing the task-specific LMs.

4. Portability issues

In order to reduce speech recognizer portability costs, it is interesting to investigate methods which limit or avoid the costly step of collecting task-specific data. Two possible solutions have been explored in this work: improving model genericity and unsupervised model adaptation. In the previous section, multi-source acoustic model training was shown to improve model genericity, obtaining recognition performance comparable to or better than the performance of task-specific acoustic models. However, this approach, which attempts to simultaneously improve performance on all target tasks, supposes the availability of manually transcribed training data.

In this section, methods are investigated to transparently adapt the BN acoustic models to a specific task using unannotated task-specific data, so as to reduce the cost of adapting to the target task. Transparent adaptation means that the procedure is automatic and can be carried out without any human expertise. The reference BN system is used to transcribe the audio data of the target task, and the recognizer hypotheses are used to adapt the acoustic models, and then both the acoustic and language models. Since the word error rate on the adaptation data is expected to be relatively high, one can think of discarding poorly recognized parts during the adaptation phase, based on confidence scores. However, recent studies have shown that only small gain is observed with data selection for acoustic model estimates as long as a large enough corpus of adaptation

data is available (Lamel et al., 2000; Wessel, 2001). Although this approach presumes that audio data have been collected, the development cost is substantially reduced since manual transcriptions are not required.

4.1. Unsupervised acoustic model adaptation

In this section, unsupervised acoustic model adaptation (Lamel et al., 2002) is investigated, making use of the unannotated task-specific data. Two adaptation procedures are explored, the first based on MAP adaptation and the second using a combination of MLLR and MAP adaptation techniques (Lefevre et al., 2001).

Cross-task unsupervised adaptation is evaluated for the three target tasks. The entire WSJ and ATIS training sets (100 and 40 h of data, respectively) were transcribed using the BN acoustic and language models. For TI-digits, the training corpus was transcribed using a mixed configuration, combining the BN acoustic models with the digit loop grammar since writing a task-specific grammar is trivial for this task and does not require collecting or transcribing any data. In each case, decoding was performed as described in Section 3.1. The resulting word error rates of the training data measured against the manual transcriptions are 1.2% for TI-digits, 25.5% for ATIS and 11.8% for WSJ. With the exception of ATIS, for which the test data appear to be "easier" than the training data, these error rates are quite close to those obtained on the evaluation sets with the same BN models (see Table 2).

Table 6 reports the word error rates obtained with the gender-dependent BN acoustic models adapted individually to each of the three tasks. The recognition tests are given for the mixed conditions, where the adapted BN acoustic models are used with the task-dependent language models (except in the first decoding pass for WSJ where the AMs were kept unchanged). The first column in Table 6 repeats the word error rates from Table 2 obtained for the mixed conditions. The second column compares two adaptation schemes: MAP alone and MLLR followed by MAP (MLLR + MAP). Unsupervised acoustic model adaptation is seen to improve performance for all tasks, except ATIS, with the best performance obtained when using the two step MLLR/MAP procedure. The relative improvements are 53% for TI-digits, 23% for read WSJ and 11% for spontaneous dictation.

Table 6
BN acoustic model adaptation – mixed conditions (task-specific LMs)

Task	Unadapted	Unsupervised		Supervised	
	BN AMs	MAP	MLLR + MAP	MAP	MLLR + MAP
TI-digits	1.7	0.8	0.8	0.5	0.5
ATIS	4.7	4.9	4.7	3.2	3.2
Read WSJ	9.0	7.3	6.9	6.7	6.5
Spontaneous dictation	13.6	12.6	11.9	11.6	11.0

Word error rates (%) for TI-digits, ATIS, read WSJ and spontaneous dictation test sets are given using task-specific lexica and LMs with: (left) BN acoustic models, (middle) unsupervised adaptation of the BN acoustic models, (right) supervised adaptation of the BN acoustic models. Two different adaptation schemes are compared: MAP alone and MLLR followed by MAP (MLLR/MAP).

In order to get an upper bound of the gain that can be expected with adaptation, the task-specific audio data and associated transcriptions were used to carry out supervised adaptation of the BN acoustic models. The resulting word error rates are given in the right column of Table 6. As expected, supervised model adaptation outperforms unsupervised adaptation for all tasks, with a substantial difference in performance for the TI-digits and ATIS tasks (32–37% relative). Smaller differences, on the order of 5%, are observed for supervised and unsupervised adaptation on the read WSJ and spontaneous dictation tasks.

The results in Table 6 show that unsupervised adaptation of the BN acoustic models improves performance when task-specific language models are used. However, these LMs make use of the manual transcripts of the acoustic training data – which are assumed to be unavailable if it is necessary to use unsupervised adaptation. Table 7 shows that an improvement in performance due to acoustic model adaptation is also observed when BN LMs are used for decoding, instead of the task-specific LMs. For the ATIS task, although a relative word error reduction of 46% is observed compared to the unadapted BN acoustic models, the word error is well above the 4.1% obtained with the task-specific system. The word error rate for the WSJ is quite close to that of the task-specific system (7.8% vs. 7.6%) and for the spontaneous dictation task the word error rate of 11.4% is even lower than the error rate with the BN reference system or with the WSJ system. The linguistic proximity of the BN and WSJ tasks largely explains these results.

These results show that better performance can be obtained by adapting somewhat generic acoustic models with task-specific data instead of directly training task-specific models. The TI-digits task is the only task for which better performance is obtained using task-dependent models rather than adapting the BN models with data from the target task. This is probably due to the limited phonemic coverage of the digit vocabulary, which is better accounted for with task-dependent phone contexts.

4.2. Unsupervised acoustic and language model adaptation

Although the experiments in the previous section showed unsupervised acoustic model adaptation to substantially reduce the word error rates, for the ATIS task there is still a large degradation in performance compared to that of a task-specific system. This performance degradation can be attributed to the linguistic mismatch between the BN and ATIS tasks. The experiments in this section investigate using the automatic transcripts for both acoustic and language model adaptation. Adapting the acoustic and language models in an incremental manner is also addressed.

Table 7
BN acoustic model adaptation – cross-task conditions (BN LMs)

Task	Unadapted	Unsupervised (MLLR + MAP)
ATIS	17.8	9.6
Read WSJ	11.6	7.8
Spontaneous dictation	12.1	11.4

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

4.2.1. One-step adaptation

The same automatic transcriptions of the 40 h of ATIS training data used for unsupervised acoustic model adaptation in the previous section, are here used for unsupervised language model adaptation. The adapted language model is obtained by a linear interpolation of the BN reference language model with a task-specific language model trained on the automatic transcriptions of the ATIS training data. To get an upper bound on the gain which can be obtained with this technique, supervised language model adaptation is also carried out. The standard BN recognizer vocabulary and pronunciations have the same OOV rate as the task-specific ATIS lexicon.

Table 8 reports word error rates using unsupervised AM and LM adaptation individually and together. Contrastive results are given with supervised adaptation (i.e., using the manual transcripts) and by training new models on the ATIS data using the automatic transcripts in place of the manual ones. The results shown in the first column are with MLLR and MAP acoustic model adaptation only (i.e., the BN LM is used for decoding). Supervised acoustic model adaptation results in a word error rate of 6.6%, which is significantly lower than the 9.6% obtained with unsupervised adaptation. It can be seen that unsupervised adaptation of the BN models outperforms the training of ATIS models on the automatic transcripts (10.8%). As a reminder, the task-specific system has a word error of 4.1% and the performance obtained under cross-task conditions without adaptation is 17.8%.

An adapted mixture LM was built by interpolating the BN 3-g backoff model with a 3-g backoff model estimated on the ATIS transcriptions (automatic transcripts in the unsupervised case and manual for the supervised one). The interpolation weight was obtained by minimizing the perplexity of a set of development texts comprised of unused prompts from the read part of ATIS0. The optimal interpolation weight for the BN LM was found to be 0.2 for both conditions. LM interpolation has nearly no effect on the perplexity: it stays the same as the task-specific model in the supervised case (16, see Table 3) and a 1 point gain is observed in the unsupervised case (25 vs. 24) with the BN LM component.

The results in the second column of Table 8 are given for language model adaptation alone, i.e., decoding with the adapted LMs and the BN acoustic models. Using the automatic transcriptions for LM adaptation results in a WER of 8.0%, which can be compared to the estimated lower bound of 4.3% obtained by using the reference manual transcriptions for adaptation. If the ATIS LM trained on the automatically transcribed data is used alone (i.e., without interpolation with

Table 8
Acoustic and language model adaptation

Training mode	Transcripts	AM	LM	AM and LM
BN → ATIS	Automatic	9.6	8.0	6.0
ATIS	Automatic	10.8	8.1	6.9
BN → ATIS	Manual	6.6	4.3	3.1

Word error rates (%) for ATIS test set with: (left) adapted BN acoustic models and the BN language model, (center) adapted BN language models and BN acoustic models, (right) adapted BN acoustic and language models. Adaptation is also compared to the training of new models (second row) and maximum possible gain estimated by using the manual transcriptions (last row). *Reminder*: task-specific system is 4.1% and cross-task condition (without adaptation) gives 17.8%.

the BN LM), the WER is 8.1%. Interpolation with the BN LM gives only a 0.1% absolute WER reduction. From these results it can be seen that the improvement obtained with language model adaptation alone is larger than that obtained with acoustic model adaptation alone, confirming the earlier conjecture that the performance gap observed in cross-task experiments for the ATIS task is mainly due to a linguistic mismatch. This is true for both unsupervised and supervised adaptations.

The rightmost column of Table 8 gives the results obtained when both the acoustic and language models are adapted. The relative error reduction observed with the combined adaptation over acoustic model adaptation alone is about 37% for unsupervised adaptation and about 53% for supervised adaptation. The gain with unsupervised adaptation is quite high given the relatively high initial error rate of the BN recognizer on the ATIS data (about 18% on the test data and 26% on the training data).

These results also show that despite the relatively low weight for the BN LM in the interpolated model, its inclusion gives a 10% relative error rate reduction compared with the ATIS LM (from 4.7% for the ATIS LM to 4.3% with the interpolated LM, using BN acoustic models). Using supervised adaptation of both the acoustic and language models yields a 25% reduction in error relative to task-specific training (4.1%).

4.2.2. Incremental adaptation

In the previous section, single step adaptation was investigated making use of an available, but potentially unannotated corpus of task-specific data. This work is extended to an incremental adaptation scheme, in which adapted models are used to annotate new untranscribed data that are in turn used for adaptation. To initiate the process, an initial set of training data is automatically transcribed using a generic transcription system. The acoustic and linguistic models of the generic system are then adapted with the 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.

This incremental scheme was applied to the ATIS task, using two adaptation steps by randomly splitting the training data into two subsets. As in the previous subsection, acoustic model adaptation is based on a combination of MLLR and MAP and language model adaptation is performed by model interpolation. The results are presented in Table 9 as a function of the amount of adaptation data. In the first step about one-third (15 h) of the ATIS training data was transcribed using the BN system, and the remaining 26 h were transcribed using the adapted acoustic and language models. Acoustic model adaptation alone reduces the initial word error rate from 17.8% to 10.5%. With language model adaptation alone using the first 15-h set of automatic transcriptions, the word error rate is reduced to 8.7%. Using all the automatic transcripts for LM adaptation results in a word error rate of 6.8%. After interpolation with the BN LM (weight is 0.2), the resulting models have a test set perplexity of 25 after the first step and 24 after the second one. Incremental adaptation of both the acoustic and language models gives a relative error reduction of almost 70–5.7%. The first 15-h subset accounts for 90% of the total reduction even though it contains only one third of the adaptation data. The second iteration gives an additional relative error reduction of 20%.

It can be seen that the performance improvement is shared between the acoustic and linguistic model adaptations. Though, it should be noted that using the additional 26 h of data for AM

Table 9
Incremental unsupervised adaptation

ATIS data		AM adaptation	LM adaptation	AM and LM adaptation
Amount	Training WER			
0	–	17.8	17.8	17.8
15 h	25.8	10.6	8.7	6.9
15 h + 26 h	13.5	10.5	6.8	5.7

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

Table 10
Variants of acoustic model adaptation

Seed models	Adaptation data	WER
BN	15 h + 26 h	10.5
Adapted BN (15 h)	26 h	9.0
BN	26 h	8.7

Word error rates (%) for ATIS after the second iteration of acoustic model adaptation for three different configuration depending of the seed models and on the adaptation data. All results are with the BN language model.

adaptation only gives a 0.1% gain, whereas the additional data substantially improves the LMs. Therefore different variations of acoustic model adaptation, after the second transcription step, were compared using the BN LMs (see Table 10). If a second adaptation is carried out by adapting the acoustic models from the first step with the additional 26 h of data, the word error rate is reduced to 9.0%. Since the word error rate of the second set of training data is much lower than the first set, the BN acoustic models were also adapted using only the 26 h data set. This results in the lowest word error rate of 8.7%, suggesting that the recognition errors in the first subset introduce inaccuracies in the acoustic models. Combining these latter acoustic models with LM adaptation gives a word error rate of 5.5%.

In these experiments, unsupervised and supervised adaptation have been investigated separately. However, in practice it is likely that a small amount of annotated training data is available along with a much larger amount of raw audio data. A first set of task-adapted models, derived from generic models in a supervised manner can be used to transcribe the raw data, which can in turn be used in another iteration of adaptation. In this context, it has been shown that the relative word error rate reduction obtained with the untranscribed part of the data increases with the amount of available data as well as with the performance of the original models (Bertoldi et al., 2003). This approach was shown to be effective in the context of a broadcast news task where less than 1 h of manually annotated data was used to bootstrap the training process (Lamel et al., 2002). In a cross-task context, a reasonable approach could be to manually correct a first set of automatic transcriptions

obtained with a generic system. This would substantially reduce the manual effort required and should lead to better performance than completely unsupervised training.

5. Conclusions

In this paper, we have explored the genericity and portability of the models used by speech recognizers, with the goal of developing more generic core technology.

The genericity of a state-of-the-art speech recognition system was assessed by testing a relatively wide-domain system on data from four tasks ranging in complexity. These models were taken from a broadcast news task, covering a wide range of acoustic and linguistic conditions. The broadcast news acoustic models were shown to be relatively task-independent since using them in place of the corresponding task-dependent acoustic models results in only a small increase in the word error rate, if a task-dependent language model was used. An exception was observed for the digit recognition task, which could be attributed to the limited phonetic coverage of this task. On a spontaneous dictation task, the broadcast news acoustic and language models were shown to be more robust to deviations in speaking style than the read–speech WSJ models.

Ways to enhance the genericity of the acoustic models by using task-specific data have been investigated. The objective here was to determine if we could improve model genericity by merging in a single model the information contained in the data from all tasks rather than developing task-specific models. Multi-source acoustic model training and adaptation were shown to improve the model accuracy, yielding recognition performance comparable to or better than that obtained with task-specific models. The various multi-source training schemes explored in the paper obtained quite comparable results with no one approach outperforming the others on every task. An improvement was also observed on the spontaneous dictation task even though no data from this task was included in the multi-source training. Multi-source acoustic training appears to be a convenient approach for designing generic models, reducing the performance gap between task-independent and task-dependent acoustic models, and for some tasks outperform task-dependent acoustic models.

Concerning portability, we have shown that unsupervised acoustic model adaptation can reduce the performance gap between task-independent and task-dependent acoustic models, and that supervised adaptation can even lead to better performance than that achieved with task-specific models.

Language model adaptation has been evaluated on the ATIS task where the linguistic mismatch with broadcast news model is the most blatant. Adapting the broadcast news acoustic and language models leads to a significant word error reduction over acoustic model adaptation alone. This is true for both supervised and unsupervised adaptation, even though the automatic transcriptions have a word error rate of about 26%. Incremental unsupervised adaptation of both the broadcast news acoustic and language models has been shown to be even more effective, giving a 70% relative error rate reduction on the ATIS task.

Acknowledgment

This work was partially financed by the European Commission under the Human Language Technology project IST-1999 11876 Coretex.

References

Bertoldi, N., Brugnara, F., Cettolo, M., Federico, M., Giuliani, D., 2003. Cross-task portability of a broadcast news speech recognition system. Speech Communication 38 (3–4), 335–347.

Dahl, D., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Pao, C., Rudnicky, A., Shriberg, E., 1994. Expanding the scope of the atis task: The atis-3 corpus. In: Proceedings of the ARPA Spoken Language Systems Technology Workshop, March, pp. 3–8.

Garofolo, J., Auzanne, C., Voorhees, E., 2000. The trec spoken document retrieval track: A success story. In: 6th RIAO Conference, Content-based Multimedia Information Access, Paris, pp. 1–20.

Gauvain, J.-L., Lee, C.-H., 1992. Map estimation of continuous density hmm: theory and application. In: DARPA Speech and Natural Language Workshop, February, pp. 185–190.

Gauvain, J.-L., Lamel, L., Adda, G., Adda-Decker, M., 1994. Speaker-independent continuous speech dictation. Speech Communication 15, 21–37.

Gauvain, J.-L., Adda, G., Lamel, L., Adda-Decker, M., 1997. Transcribing broadcast news: the limsi nov96 hub4 system. In: Proceedings of the ARPA Spoken Language Systems Technology Workshop, Febuary, pp. 56–63.

Gauvain, J.-L., Lamel, L., de Kercadio, Y., Adda, G., 2000. Transcription and indexation of broadcast data. In: Proceedings of the IEEE ICASSP, Istanbul, pp. 1663–1666.

Graff, D., 1997. The 1996 broadcast new speech and language-model corpus. In: 1997 DARPA Speech Recognition Workshop, Chantilly.

Kemp, T., Waibel, A., 1999. Unsupervised training of a speech recognizer: recent experiments. In: Proceedings of the ESCA Eurospeech, vol. 6, Budapest, pp. 2725–2728.

Kubala, F., Bellegarda, J., Cohen, J., Pallett, D., Paul, D., Phillips, M., Rajasekaran, R., Richardson, F., Riley, M., Rosenfeld, R., Roth, B., Weintraub, M., 1994. The hub and spoke paradigm for csr evaluation. In: Proceedings of the ARPA Spoken Language Systems Technology Workshop, pp. 9–14.

Lamel, L., Gauvain, J.-L., Adda, G., 2002. Lightly supervised and unsupervised acoustic model training. Computer Speech and Language 16 (1), 115–129.

Lefevre, F., Gauvain, J.-L., Lamel, L, 2001. Improving genericity for task-independent speech recognition. In: Proceedings of the ISCA Eurospeech, vol. 2, Aalborg, pp. 1241–1244.

Lamel, L., Gauvain, J.L., Adda, G, 2000. Lightly supervised acoustic model training. In: Proceedings of the ISCA ITRW ASR2000, Paris, pp. 150–154.

Lefevre, F., Gauvain, J.-L., Lamel, L., 2001. Toward task-independent speech recognition. In Proceedings of the IEEE ICASSP, vol. I, Salt Lake City, pp. 521–524.

Leggetter, C., Woodland, P., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer Speech and Language 9, 171–185.

Leonard, R., 1984. A database for speaker-independent digit recognition. In: Proceedings of the ICASSP, vol. III, pp. 42.11–42.14.

Normandin, Y., Cardin, R., de Mori, R., 1994. High-performance connected digit recognition using maximum mutual information estimation. IEEE Transactions on Speech and Audio Processing 2 (2), 299–311.

Pallett, D.S., Fiscus, J.G., Garofolo, J., Martin, A., Przybocki, M., 1999. 1998 broadcast news benchmark test results. In: Proceedings of the DARPA Broadcast News Workshop, pp. 5–12.

Paul, D., Baker, J., 1992. The design for the wall street journal-based csr corpus. In: Proceedings of the ICSLP, Banff, pp. 899–902.

Price, P., 1990. Evaluation of spoken language systems: the atis domain. In: Proceedings of the DARPA Workshop on Speech and Natural Language, Hidden Valley, PA.

Schultz, T., Waibel, A., 2001. Language independent and language adaptive acoustic modeling for speech recognition. Speech Communication 35 (1–2), 31–51.

Wessel, F., 2001. Unsupervised training of acoustic models for large vocabulary speech recognition. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, CDROM, IEEE Catalog No. 01EX544, Madonna di Campiglio.

Woodland, P., Leggetter, C., Odell, J., Valtchev, V., Young, S., 1995. The 1994 htk large vocabulary speech recognition system. In: Proceedings IEEE ICASSP, vol. I, Detroit, pp. 73–76.

Zavaliagkos, G., Colthurst, T., 1998. Utilizing untranscribed training data to improve performance. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 301–305.

Zavaliagkos, G., Anastsakos. T., Chou. G., Lapre, C., Kubala, F., Makhoul, J., Nguyen, L., Schwartz, R., Zhao, Y., 1994. Improved search, acoustic, and language modeling in the bbn byblos large vocabulary csr systems. In: Proceedings of the ARPA Spoken Language Systems Technology Workshop, pp. 81–88.