Portability Issues for Speech Recognition Technologies*

Lori Lamel, Fabrice Lefevre, Jean-Luc Gauvain and Gilles Adda

Spoken Language Processing Group,
CNRS-LIMSI, 91403 Orsay, France
{lamel,lefevre,gauvain,gadda}@limsi.fr

ABSTRACT

Although there has been regular improvement in speech recognition technology over the past decade, speech recognition is far from being a solved problem. Most recognition systems are tuned to a particular task and porting the system to a new task (or language) still requires substantial investment of time and money, as well as expertise. Todays state-of-the-art systems rely on the availability of large amounts of manually transcribed data for acoustic model training and large normalized text corpora for language model training. Obtaining such data is both time-consuming and expensive, requiring trained human annotators with substantial amounts of supervision.

In this paper we address issues in speech recognizer portability and activities aimed at developing generic core speech recognition technology, in order to reduce the manual effort required for system development. Three main axes are pursued: assessing the genericity of wide domain models by evaluating performance under several tasks; investigating techniques for lightly supervised acoustic model training; and exploring transparent methods for adapting generic models to a specific task so as to achieve a higher degree of genericity.

1. INTRODUCTION

The last decade has seen impressive advances in the capability and performance of speech recognizers. Todays state-of-the-art systems are able to transcribe unrestricted continuous speech from broadcast data with acceptable performance. The advances arise from the increased accuracy and complexity of the models, which are closely related to the availability of large spoken and text corpora for training, and the wide availability of faster and cheaper computational means which have enabled the development and implementation of better training and decoding algorithms. Despite the extent of progress over the recent years, recognition accuracy is still extremely sensitive to the environmental conditions and speaking style: channel quality, speaker characteristics, and background noise have an important impact on the acoustic component of the speech recognizer, whereas the speaking style and the discourse domain have a large impact on the linguistic component.

In the context of the EC IST-1999 11876 project CORETEX we are investigating methods for fast system development, as well as development of systems with high genericity and adaptability. By fast system development we refer to: language support, i.e., the capability of porting technology to different languages at a reasonable cost; and task portability, i.e. the capability to easily adapt a technology to a new task by exploiting limited amounts of domain-specific knowledge. 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. Concerning the acoustic modeling component, genericity implies that it is robust to the type and bandwidth of the channel, the acoustic environment, the speaker type and the speaking style. Unsupervised normalization and adaptation techniques evidently should be used to enhance performance further when the system is exposed to data of a particular type.

With today's technology, the adaptation of a recognition system to a new task or new language requires the availability of sufficient amount of transcribed training data. When changing to new domains, usually no exact transcriptions of acoustic data are available, and the generation of such transcribed data is an expensive process in terms of manpower and time. On the other hand, there often exist incomplete information such as approximate transcriptions, summaries or at least key words, which can be used to provide supervision in what can be referred to as "informed speech recognition". Depending on the level of completeness, this information can be used to develop confidence measures with adapted or trigger language models or by approximate alignments to automatic transcriptions. Another approach is to use existing recognizer components (developed for other tasks or languages) to automatically transcribe task-specific training data. Although in the beginning the error rate on new data is likely to be rather high, this speech data can be used to re-train a recognition system. If carried out in an iterative manner, the speech data base for the new domain can be cumulatively extended over time *without* direct manual transcription.

The overall objective of the work presented here is to reduce the speech recognition development cost. One aspect is to develop "generic" core speech recognition technology, where by "generic" we mean a transcription engine that will work reasonably well on a wide range of speech transcription tasks, ranging from digit recognition to large vocabulary conversational telephony speech, without the need for costly task-specific training data. To start with we assess the genericity of wide domain models under cross-task con-

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

Table 1: Brief descriptions and best reported error rates for the corpora used in this work.

Corpus	*Test Year*	*Task*	*Train (#spkr)*	*Test (#spkr)*	*Textual Resources*	*Best WER*
BN	98	TV & Radio News	200h	3h	Closed-captions, commercial transcripts, manual transcripts of audio data	13.5
TI-digits	93	Small Vocabulary	3.5h (112)	4h (113)	-	0.2
ATIS	93	H-M Dialog	40h (137)	5h (24)	Transcriptions	2.5
WSJ	95	News Dictation	100h (355)	45mn (20)	Newspaper, newswire	6.6
S9_WSJ	93	Spontaneous Dictation		43mn (10)	Newspaper, newswire	19.1

ditions, i.e., by recognizing task-specific data with a recognizer developed for 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), read and spontaneous text dictation (WSJ), and goal-oriented spoken dialog (ATIS). The broadcast news task is quite general, covering a wide variety of linguistic and acoustic events in the language, ensuring reasonable coverage of the target task. In addition, there are sufficient acoustic and linguistic training data available for this task that accurate models covering a wide range of speaker and language characteristics can be estimated.

Another research area is the investigation of lightly supervised techniques for acoustic model training. The strategy taken is to use a speech recognizer to transcribe unannotated data, which are then used to estimate more accurate acoustic models. The light supervision is applied to the broadcast news task, where unlimited amounts of acoustic training data are potentially available. Finally we apply the lightly supervised training idea as a transparent method for adapting the generic models to a specific task, thus achieving a higher degree of genericity. In this work we focus on reducing training costs and task portability, and do not address language transfer.

We selected the LIMSI broadcast news (BN) transcription system as the generic reference system. The BN task covers a large number of different acoustic and linguistic situations: planned to spontaneous speech; native and non-native speakers with different accents; close-talking microphones and telephone channels; quiet studio, on-site reports in noisy places to musical background; and a variety of topics. In addition, a lot of training resources are available including a large corpus of annotated audio data and a huge amount of raw audio data for the acoustic modeling; and large collections of closed-captions, commercial transcripts, newspapers and newswires texts for linguistic modeling. The next section provides an overview of the LIMSI broadcast news transcription system used as our generic system.

2. SYSTEM DESCRIPTION

The LIMSI broadcast news transcription system has two main components, the audio partitioner and the word recognizer. Data partitioning [6] serves to divide the continuous audio stream into homogeneous segments, associating appropriate labels for cluster, gender and bandwidth with the segments. The speech recognizer 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 [13] 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 Hub4 Broadcast News corpus (the LDC 1996 and 1997 Broadcast News Speech collections) [9]. Gender-dependent acoustic models were built using MAP adaptation of SI seed models for wide-band and telephone band speech [7]. The models contain 28000 position-dependent, cross-word triphone models with 11700 tied states and approximately 360k Gaussians [8].

The baseline language models are obtained by interpolation of models trained on 3 different data sets (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 recognition vocabulary contains 65120 words and has a lexical coverage of over 99% on all evaluation test sets from the years 1996-1999. A pronunciation graph is associated with each word so as to allow for alternate pronunciations. The pronunciations make use of a set of 48 phones set, where 3 phone units represent silence, filler words, and breath noises. The lexicon contains compound words for about 300 frequent word sequences, as well as word entries for common acronyms, providing an easy way to allow for reduced pronunciations [6].

The LIMSI 10x system obtained a word error of 17.1% on the 1999 DARPA/NIST evaluation set and can transcribe unrestricted broadcast data with a word error of about 20% [8].

3. TASK INDEPENDENCE

Our first step in developing a "generic" speech transcription engine is to assess the most generic system we have under cross-task conditions, i.e., by recognizing task-specific data with a recognizer developed for a different task. Three representative tasks have been retained as target tasks: small vocabulary recognition (TI-digits), goal-oriented human-machine spoken dialog (ATIS), and dictation of texts (WSJ). The broadcast news transcription task (Hub4E) serves as the baseline. The main criteria for the task selection were that they are realistic enough and task-specific data should be available. The characteristics of these four tasks and the available corpora are summarized in Table 1.

For the small vocabulary recognition task, experiments are carried out on the adult speaker portion of the TI-digits corpus [14], containing over 17k utterances from a total of 225 speakers. The vocabulary contains 11 words, the digits '1' to '9', plus 'zero' and 'oh'. Each speaker uttered two versions of each digit in isolation and 55 digit strings. The database is divided into training and test sets (roughly 3.5 hours each, corresponding to 9k strings). The speech is of high quality, having been collected in a quiet environment. The best reported WERs on this task are around 0.2-0.3%. The digit phonemic coverage being very low, only 108 context-dependent models are used in our recognition system. The task-

Table 2: Word error rates (%) for BN98, TI-digits, ATIS94, WSJ95 and S9_WSJ93 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.

Test Set	*BN models*	*Task LMs*	*Task models*
BN98	13.6	13.6	13.6
TI-digits	17.5	1.7	0.4
ATIS94	22.7	4.7	4.4
WSJ95	11.6	9.0	7.6
S9_WSJ93	12.1	13.6	15.3

specific LM for the TI-digits is a simple grammar allowing any sequence of up to 7 digits. Our task-dependent system performance is 0.4% WER.

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 (SPREC) ATIS-3 data (*ATIS94*) [4] is used for testing purposes. The test data amounts for nearly 5 hours of speech from 24 speakers recorded with a close-talking microphone. Around 40h of speech data are available for training. 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 acoustic models used in our task-specific system include 1641 context-dependent phones with 4k independent HMM states. A back-off trigram language model has been 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.). The WER obtained with our task-dependent system is 4.4%.

For the dictation task, the *Wall Street Journal* continuous speech recognition corpus [17] is used, abiding by the ARPA 1995 Hub3 test (*WSJ95*) 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. The best result reported at the time of the evaluation was 6.6%. A contrastive experiment is carried out with the WSJ93 Spoke 9 data comprised of 200 spontaneous sentences spoken by journalists [11]. The best performance reported in the 1993 evaluation on the spontaneous data was 19.1% [18], however lower word error rates have since been reported on comparable test sets (14.1% on the WSJ94 Spoke 9 test data). 21000 context and position-dependent models have been trained for the WSJ system, with 9k independent HMM states. A 65k-word vocabulary was selected and a back-off trigram model obtained by interpolating models trained on different data sets (training utterance transcriptions and newspapers data). The task-dependent WSJ system has a WER of 7.6% on the read speech test data and 15.3% on the spontaneous data.

For the BN transcription task, we follow the conditions of the 1998 ARPA Hub4E evaluation (*BN98*) [15]. The acoustic training data is comprised of 150 hours of North-American TV and radio shows. The best overall result on the 1998 baseline test was 13.5%.

Three sets of experiments are reported. The first are cross-task recognition experiments carried out using the BN acoustic and language models to decode the test data for the other tasks. The second set of experiments made use of mixed models, that is the BN acoustic models and task-specific LMs. Due to the different evaluation paradigms, some minor modifications were made in the transcription procedure. First of all, in contrast with the BN data, the data for the 3 tasks is already segmented into individual utterances so the partitioning step was eliminated. With this exception, the decoding process for the WSJ task is exactly the same as described in the previous section. For the TI-digits and ATIS tasks, word decoding is carried out in a single trigram pass, and no speaker adaptation was performed.

The WERs obtained for the three recognition experiments are reported in Table 2. A comparison with Table 1 shows that the performances of the task-dependent models are close to the best reported results even though we did not devote too much effort in optimizing these models. We can also observe by comparing the task-dependent (Table 2, right) and mixed (Table 2, middle) conditions, that the BN acoustic models are relatively generic. These models seem to be a good start towards truly task-independent acoustic models. By using task-specific language models For the TI-digits and ATIS we can see that the gap in performance is mainly due a linguistic mismatch. For WSJ the language models are more closely matched to BN and only a small 1.6% WER reduction is obtained. On the spontaneous journalist dictation (WSJ S9 spoke) 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.

Prior to introducing our approach for lightly supervised acoustic model training, we describe our standard training procedure in the next section.

4. ACOUSTIC MODEL TRAINING

HMM training requires an alignment between the audio signal and the phone models, which usually relies on a perfect orthographic transcription of the speech data and a good phonetic lexicon. In general it is easier to deal with relatively short speech segments so that transcription errors will not propagate and jeopardize the alignment. The orthographic transcription is usually considered as ground truth and training is done in a closely supervised manner. For each speech segment the training algorithm is provided with the exact orthographic transcription of what was spoken, i.e., the word sequence that the speech recognizer should hypothesize when confronted with the same speech segment.

Training acoustic models for a new corpus (which could also reflect a change of task and/or language), usually entails the following sequence of operations once the audio data and transcription files have been loaded:

1. Normalize the transcriptions to a common format (some adjustment is always needed as different text sources make use of different conventions).
2. Produce a word list from the transcriptions and correct blatant errors (these include typographical errors and inconsistencies).
3. Produce a phonemic transcription for all words not in our master lexicon (these are manually verified).
4. Align the orthographic transcriptions with the signal using existing models and the pronunciation lexicon (or bootstrap models from another task or language). This procedure often rejects a substantial portion of the data, particularly for long segments.
5. Eventually correct transcription errors and realign (or just ignore these if enough audio data is available)
6. Run the standard EM training procedure.

This sequence of operations is usually iterated several times to refine the acoustic models. In general each iteration recovers a portion of the rejected data.

5. LIGHTLY SUPERVISED ACOUSTIC MODEL TRAINING

One can imagine training acoustic models in a less supervised manner, by using an iterative procedure where instead of using manual transcriptions for alignment, at each iteration the most likely word transcription given the current models and all the information available about the audio sample is used. This approach still fits within the EM training framework, which is well-suited for missing data training problems. A completely unsupervised training procedure is to use the current best models to produce an orthographic transcription of the training data, keeping only words that have a high confidence measure. Such an approach, while very enticing, is limited since the only supervision is provided by the confidence measure estimator. This estimator must in turn be trained on development data, which needs to be small to keep the approach interesting.

Between using carefully annotated data such as the detailed transcriptions provided by the LDC and no transcription at all, there is a wide spectrum of possibilities. What is really important is the cost of producing the associated annotations. Detailed annotation requires on the order of 20-40 times real-time of manual effort, and even after manual verification the final transcriptions are not exempt from errors [2]. Orthographic transcriptions such as closed-captions can be done in a few times real-time, and therefore are quite a bit less costly. These transcriptions have the advantage that they are already available for some television channels, and therefore do not have to be produced specifically for training speech recognizers. However, closed-captions are a close, but not exact transcription of what is being spoken, and are only coarsely time-aligned with the audio signal. Hesitations and repetitions are not marked and there may be word insertions, deletions and changes in the word order. They also are missing some of the additional information provided in the detailed speech transcriptions such as the indication of acoustic conditions, speaker turns, speaker identities and gender and the annotation of non-speech segments such as music. NIST found the disagreement between the closed-captions and manual transcripts on a 10 hour subset of the TDT-2 data used for the SDR evaluation to be on the order of 12% [5].

Another approach is to make use of other possible sources of contemporaneous texts from newspapers, newswires, summaries and the Internet. However, since these sources have only an indirect correspondence with the audio data, they provide less supervision.

The basic idea is of light supervision is to use a speech recognizer to automatically transcribe unannotated data, thus generating "approximate" labeled training data. By iteratively increasing the amount of training data, more accurate acoustic models are obtained, which can then be used to transcribe another set of unannotated data. The modified training procedure used in this work is:

1. Train a language model on all texts and closed captions after normalization
2. Partition each show into homogeneous segments and label the acoustic attributes (speaker, gender, bandwidth) [6]
3. Train acoustic models on a very small amount of manually annotated data (1h)
4. Automatically transcribe a large amount of training data
5. (Optional) Align the closed-captions and the automatic transcriptions (using a standard dynamic programming algorithm)
6. Run the standard acoustic model training procedure on the speech segments (in the case of alignment with the closed captions only keep segments where the two transcripts are in agreement)
7. Reiterate from step 4.

It is easy to see that the manual work is considerably reduced, not only in generating the annotated corpus but also during the training procedure, since we no longer need to extend the pronunciation lexicon to cover all words and word fragments occurring in the training data and we do not need to correct transcription errors. This basic idea was used to train acoustic models using the automatically generated word transcriptions of the 500 hours of audio broadcasts used in the spoken document retrieval task (part of the DARPA TDT-2 corpus used in the SDR'99 and SDR'00 evaluations) [3]. This corpus is comprised of 902 shows from 6 sources broadcast between January and June 1998: CNN Headline News (550 30-minute shows), ABC World News Tonight (139 30-minute shows), Public Radio International The World (122 1-hour shows), Voice of America VOA Today and World Report (111 1-hour shows). These shows contain about 22k stories with time-codes identifying the beginning and end of each story.

First, the recognition performance as a function of the available acoustic and language model training data was assessed. Then we investigated the accuracy of the acoustic models obtained after recognizing the audio data using different levels of supervision via the language model. With the exception of the baseline Hub4 language models, none of the language models include a component estimated on the transcriptions of the Hub4 acoustic training data. The language model training texts come from contemporaneous sources such as newspapers and newswires, and commercial summaries and transcripts, and closed-captions. The former sources have only an indirect correspondence with the audio data and provide less supervision than the closed captions. For each set of LM training texts, a new word list was selected based on the word frequencies in the training data. All language models are formed by interpolating individual LMs built on each text source. The interpolation coefficients were chosen in order to minimize the perplexity on a development set composed of the second set of the Nov98 evaluation data (3h) and a 2h portion of the TDT2 data from Jun98 (not included in the LM training data). The following combinations were investigated:

- **LMa** (baseline Hub4 LM): newspaper+newswire (NEWS), commercial transcripts (COM) predating Jun98, acoustic transcripts
- **LMn_t_c:** NEWS, COM, closed-captions through May98
- **LMn_t:** NEWS, COM through May98
- **LMn_c:** NEWS, closed-captions through May98
- **LMn:** NEWS through May98
- **LMn_to:** NEWS through May98, COM through Dec97
- **LMno:** NEWS through Dec97

Table 3: Word error rate for various conditions using acoustic models trained on the HUB4 training data with detailed manual transcriptions. All runs were done in less than 10xRT, except the last row. "1S" designates one set of gender-independent acoustic models, whereas "4S" designates four sets of gender and bandwidth dependent acoustic models.

Training	*Conditions*	*bn99_1*	*bn99_2*	*Average*
1h	1S, LMn_t_c	35.2	31.9	33.3
69h	1S, LMn_t_c	20.2	18.0	18.9
123h	1S, LMn_t_c	19.3	17.1	18.0
123h	4S, LMn_t_c	18.5	16.1	17.1
123h	4S, LMa	18.3	16.3	17.1
123h	4S, LMa, 50x	17.1	14.5	15.6

Table 4: Word error rate for different language models and increasing quantities of automatically labeled training data on the 1999 evaluation test sets using gender and bandwidth independent acoustic models. LMn_t_c: NEWS, COM, closed-captions through May98 LMn_t: NEWS, COM through May98 LMn_c: NEWS, closed-captions through May98 LMn: NEWS through May98 LMn_to: NEWS through May98, COM through Dec97 LMno: NEWS through Dec97.

Amount of training data		*%WER*					
raw	*unfiltered*	*LMn_t_c*	*LMn_t*	*LMn_c*	*LMn*	*LMn_to*	*LMno*
150h	123h	18.0	18.6	19.1	20.6	18.7	20.9
1h	1h	33.3	33.7	34.4	35.9	33.9	36.1
14h	8h	26.4	27.6	27.4	29.0	27.6	30.6
28h	17h	25.2	25.7	25.6	28.1	25.7	28.9
58h	28h	24.3	25.2	25.7	27.4	25.1	27.9

It should be noted that all of the conditions include newspaper and newswire texts from the same epoch as the audio data. These provide an important source of knowledge particularly with respect to the vocabulary items. Conditions which include the closed captions in the LM training data provide additional supervision in the decoding process when transcribing audio data from the same epoch.

For testing purposes we use the 1999 Hub4 evaluation data, which is comprised of two 90 minute data sets selected by NIST. The first set was extracted from 10 hours of data broadcast in June 1998, and the second set from a set of broadcasts recorded in August-September 1998 [16]. All recognition runs were carried out in under 10xRT unless stated otherwise. The LIMSI 10x system obtained a word error of 17.1% on the evaluation set (the combined scores in the penultimate row in Table 3 4S, LMa) [8]. The word error can be reduced to 15.6% for a system running at 50xRT (last entry in Table 3).

As can be seen in Table 3, the word error rates with our original Hub4 language model (LMa) and the one without the transcriptions of the acoustic data (LMn_t_c) give comparable results using the 1999 acoustic models trained on 123 hours of manually annotated data (123h, 4S). The quality of the different language models listed above are compared in the first row of Table 3 using speaker-independent (1S) acoustic models trained on the same Hub4 data (123h). As can be observed, removing any text source leads to a degradation in recognition performance. It appears it is more important to include commercial transcripts (LMn_t), even if they are old (LMn_to) than the closed captions (LMn_c). This suggests that the commercial transcripts more accurately represent spoken language than closed-captioning. Even if only newspaper and newswire texts are available (LMn), the word error increases by only 14% over the best configuration (LMn_t_c), and even using older newspaper and newswire texts (LMno) does not substantially increase the word error rate. The second row of Table 3 gives the word error rates with acoustic models trained on only 1 hour of manually transcribed data. These are the models used to initialize the process of automatically transcribing large quantities of data. These word error rates range from 33% to 36% across the language models.

We compared a straightforward approach of training on all the automatically annotated data with one in which the closed-captions are used to filter the hypothesized transcriptions, removing words that are "incorrect". In the filtered case, the hypothesized transcriptions are aligned with the closed captions story by story, and only regions where the automatic transcripts agreed with the closed captions were kept for training purposes. To our surprise, somewhat comparable recognition results were obtained both with and without filtering, suggesting that inclusion of the closed-captions in the language model training material provided sufficient supervision (see Table 5).[1] It should be noted that in both cases the closed-caption story boundaries are used to delimit the audio segments after automatic transcription.

To investigate this further we are assessing the effects of reducing the amount of supervision provided by the language model training texts on the acoustic model accuracy (see Table 4). With 14 hours (raw) of approximately labeled training data, the word error is reduced by about 20% for all LMs compared with training on 1h of data which has carefully manual transcriptions. Using larger amounts of data transcribed with the same initial acoustic models gives smaller improvements, as seen by the entries for 28h and 58h. The commercial transcripts (LMn+t and LMn+to), even if predating the data epoch, are seen to be more important than the closed-captions (LMn+c), supporting the earlier observation that they are closer to spoken language. Even if only news texts from the same period (LMn) are available, these provide adequate supervision for lightly supervised acoustic model training.

Table 5: Word error rates for increasing quantities of automatically label training data on the 1999 evaluation test sets using gender and bandwidth independent acoustic models with the language model LMn_t_c (trained on NEWS, COM, closed-captions through May98).

Amount of training data			*%WER*	
raw	*unfiltered*	*filtered*	*unfiltered*	*filtered*
14h	8h	6h	26.4	25.7
28h	17h	13h	25.2	23.7
58h	28h	21h	24.3	22.5
140h	76h	57h	22.4	21.1
287h	140h	108h	21.0	19.9
503h	238h	188h	20.2	19.4

6. TASK ADAPTATION

The experiments reported in the section 3 show that while direct recognition with the reference BN acoustic models gives relatively

[1] The difference in the amounts of data transcribed and actually used for training is due to three factors. The first is that the total duration includes non-speech segments which are eliminated prior to recognition during partitioning. Secondly, the story boundaries in the closed captions are used to eliminate irrelevant portions, such as commercials. Thirdly, since there are many remaining silence frames, only a portion of these are retained for training.

Table 6: Word error rates (%) for TI-digits, ATIS94, WSJ95 and S9_WSJ93 test sets after recognition with three different configurations, all including task-specific lexica and LMs: (left) BN acoustic models, (middle left) unsupervised adaptation of the BN acoustic models, (middle right) supervised adaptation of the BN acoustic models and (right) task-dependent acoustic models.

Test Set	*BN models*	*Unsupervised Adaptation BN models*	*Supervised Adaptation BN models*	*Task-dep. models*
TI-digits	1.7	0.8	0.5	0.4
ATIS94	4.7	4.7	3.2	4.4
WSJ95	9.0	6.9	6.7	7.6
S9_WSJ93	13.6	12.6	11.4	15.3

competitive results, the WER on the targeted tasks can still be improved. Since we want to minimize the cost and effort involved in tuning to a target task, we are investigating methods to transparently adapt the reference acoustic models. By transparent we mean that the procedure is automatic and can be carried out without any human expertise. We therefore apply the approach presented in the previous section, that is the reference BN system is used to transcribe the training data of the destination task. This supposes of course that audio data have been collected. However, this can be carried out with an operational system and the cost of collecting task-specific training data is 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. Thus there is no need to design a new set of models based on the training data characteristics. Adaptation is also preferred to the training of new models as it is likely that the new training data will have a lower phonemic contextual coverage than the original reference models.

The cross-task unsupervised adaptation is evaluated for the tasks: TI-digits, ATIS and WSJ. The 100 hours of the WSJ data were transcribed using the BN acoustic and language models. For ATIS, only 26 of the 40 hours of training data from 276 speakers were transcribed, due to time constraints. For TI-digits, the training data was transcribed using a mixed configuration, combining the BN acoustic models with the simple digit loop grammar.[2] For completeness we also used the task-specific audio data and the associated transcriptions to carry out supervised adaptation of the BN models.

Gender-dependent acoustic models were estimated using the corresponding gender-dependent BN models as seeds and the gender-specific training utterances as adaptation data. For WSJ and ATIS, the speaker ids were directly used for gender identification since in previous experiments with this test set there were no gender classification errors. Only the acoustic models used in the second and third word decoding passes have been adapted. For the TI-digits, the gender of each training utterance was automatically classified by decoding each utterance twice, once with each set of gender-dependent models. Then, the utterance gender was determined based on the best global score between the male and female models (99.0% correct classification).

Both the MLLR and MAP adaptation techniques were applied. The recognition tests were carried out under mixed conditions (i.e., with the adapted acoustic models and the task-dependent LM). The BN models are first adapted using MLLR with a global transformation, followed by MAP adaptation.

The word error rates obtained with the task-adapted BN models are given in Table 6 for the four test sets. Using unsupervised adaptation the performance is improved for TIdigits (53% relative), WSJ (19% relative) and S9 (7% relative).

The manual transcriptions for the targeted tasks were used to carry out supervised model adaptation. The results (see the 4th column of Table 6) show a clear improvement over unsupervised adaptation for both the TI-digits (60% relative) and ATIS (47% relative) tasks. A smaller gain of about 10% relative is obtained for the spontaneous dictation task, and only 3% relative for read WSJ data. The gain appears to be correlated with the WER of the transcribed data: the difference between BN and task specific models is smaller for WSJ than ATIS and TI-digits. The TI-digit task is the only task for which the best performance is obtained using task-dependent models rather than BN models adapted with supervised. For the other tasks, the lowest WER is obtained when the supervised adapted BN acoustic models are used: 3.2% for ATIS, 6.7% for WSJ and 11.4% for S9. This result confirms our hypothesis that better performance can be achieved by adapting generic models with task-specific data than by directly training task-specific models.

7. CONCLUSIONS

This paper has explored methods to reduce the cost of developing models for speech recognizers. Two main axes have been explored: developing generic acoustic models and the use of low cost data for acoustic model training.

We have explored the genericity of state-of-the-art speech recognition systems, by testing a relatively wide-domain system on data from three tasks ranging in complexity. The generic models were taken from the broadcast news task which covers a wide range of acoustic and linguistic conditions. These acoustic models are 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. There remains a large difference in performance on the digit recognition task which can be attributed to the limited phonetic coverage of this task. On a spontaneous WSJ dictation task, the broadcast news acoustic and language are more robust to deviations in speaking style than the read-speech WSJ models. We also have shown 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 these may be a special case.

We have investigated the use of low cost data to train acoustic models for broadcast news transcription, with supervision provided

[2] In order to assess the quality of the automatic transcription, we compared the system hypotheses to the manually provided training transcriptions. For resulting word error rates on the training data are 11.8% for WSJ, 29.1% for ATIS and 1.2% for TI-digits.

the language models. Recognition results obtained with acoustic models trained on large quantities of automatically annotated data are comparable (under a 10% relative increase in word error) to results obtained with acoustic models trained on large quantities of manually annotated data. Given the significantly higher cost of detailed manual transcription (substantially more time consuming than producing commercial transcripts, and more expensive since closed captions and commercial transcripts are produced for other purposes), such approaches are very promising as they require substantial computation time, but little manual effort. Another advantage offered by this approach is that there is no need to extend the pronunciation lexicon to cover all words and word fragments occurring in the training data. By eliminating the need for manual transcription, automated training can be applied to essentially unlimited quantities of task-specific training data. While the focus of our work has been on reducing training costs and task portability, we have been exploring these in a multi-lingual context.

REFERENCES

[1] G. Adda, M. Jardino, J.L. Gauvain, "Language Modeling for Broadcast News Transcription," *ESCA Eurospeech'99*, Budapest, **4**, pp. 1759-1760, Sept. 1999.

[2] C. Barras, E. Geoffrois et al.,"Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, **33**(1-2), pp. 5-22, Jan. 2001.

[3] C. Cieri, D. Graff, M. Liberman, "The TDT-2 Text and Speech Corpus," *DARPA Broadcast News Workshop*, Herndon. (see also http://morph.ldc.upenn.edu/TDT).

[4] D. Dahl, M. Bates *et al.*, "Expanding the Scope of the ATIS Task : The ATIS-3 Corpus," *Proc. ARPA Spoken Language Systems Technology Workshop*, Plainsboro, NJ, pp. 3-8, 1994.

[5] J. Garofolo, C. Auzanne, E. Voorhees, W. Fisher, "1999 TREC-8 Spoken Document Retrieval Track Overview and Results," *8th Text Retrieval Conference TREC-8*, Nov. 1999.

[6] J.L. Gauvain, G. Adda, *et al.*, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *Proc. ARPA Speech Recognition Workshop*, pp. 56-63, Chantilly, Feb. 1997.

[7] J.L. Gauvain, C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observation of Markov Chains," *IEEE Trans. on SAP*, **2**(2), pp. 291-298, April 1994.

[8] J.L. Gauvain, L. Lamel, "Fast Decoding for Indexation of Broadcast Data," *ICSLP'2000*, **3**, pp. 794-798, Beijing, Oct. 2000.

[9] D. Graff, "The 1996 Broadcast News Speech and Language-Model Corpus," *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, pp. 11-14, Feb. 1999.

[10] T. Kemp, A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *Eurospeech'99*, **6**, Budapest, pp. 2725-2728, Sept. 1999.

[11] F. Kubala, J. Cohen *et al.*, "The Hub and Spoke Paradigm for CSR Evaluation," *Proc. ARPA Spoken Language Systems Technology Workshop*, Plainsboro, NJ, pp. 9-14, 1994.

[12] L. Lamel, J.L. Gauvain, G. Adda, "Lightly Supervised Acoustic Model Training," *Proc. ISCA ITRW ASR2000*, pp. 150-154, Paris, Sept. 2000.

[13] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2), pp. 171-185, 1995.

[14] R.G. Leonard, "A Database for speaker-independent digit recognition," *Proc. ICASSP*, 1984.

[15] D.S. Pallett, J.G. Fiscus, *et al.* "1998 Broadcast News Benchmark Test Results," *Proc. DARPA Broadcast News Workshop*, pp. 5-12, Herndon, VA, Feb. 1999.

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

[17] D.B. Paul, J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *Proc. ICSLP*, Kobe, Nov. 1992.

[18] G. Zavaliagkos, T. Anastsakos *et al.*, "Improved Search, Acoustic, and Language Modeling in the BBN BYBLOS Large Vocabulary CSR Systems," *Proc. ARPA Spoken Language Systems Technology Workshop*, Plainsboro, NJ, pp. 81-88, 1994.

[19] G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, pp. 301-305, Feb. 1998.