Improving Genericity for Task-Independent Speech Recognition*

Fabrice Lefevre, Jean-Luc Gauvain and Lori Lamel

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

10.21437/Eurospeech.2001-322

ABSTRACT

Although there have been regular improvements in speech recognition technology over the past decade, speech recognition is far from being a solved problem. Recognition systems are usually tuned to a particular task and porting the system to a new task (or language) is both time-consuming and expensive. In this paper, issues in speech recognizer portability are addressed through the development of generic core speech recognition technology. First, the genericity of wide domain models is assessed by evaluating performance on several tasks. Then, the use of transparent methods for adapting generic models to a specific task is explored. Finally, further techniques are evaluated aiming at enhancing the genericity of the wide domain models. We show that unsupervised acoustic model adaptation and multi-source training can reduce the performance gap between task-independent and task-dependent acoustic models, and for some tasks even out-perform task-dependent acoustic models.

1. INTRODUCTION

In the context of the EC IST-1999 CORETEX project we are investigating methods for development of systems with high degree of genericity and adaptability. Genericity and adaptability refer to the capacity of the technology to work properly on a wide range of tasks and to dynamically keep models up to date using contemporary data. The more robust the initial generic system is, the less there is a need for adaptation. Unsupervised normalization and adaptation techniques evidently should be used to further enhance performance when the system is exposed to data of a particular type.

The main objective of this work is to develop *generic* core speech recognition technology. By generic we mean a transcription engine that will work reasonably well on a wide range of speech tasks, ranging from digit recognition to large vocabulary conversational telephony speech, without the need for costly task-specific training data.

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 the acoustic data are available, and the generation of such transcribed data is an expensive process in terms of manpower and time. An approach is to use existing recognizer components (developed for other tasks or languages) to automatically transcribe the task-specific training data. These data can in turn be used to adapt the initial models to the new task. A step beyond is to use the task-specific training data from multiple sources to enhance the genericity of the reference models. To do so, a variety of approaches are possible: pooling data, interpolating models or via single or multi-step model adaptation. The objective here is to obtain results with the new generic models comparable or better than the respective task-dependent results for all tasks under consideration.

To start with we assess the genericity of wide domain models under cross-task conditions, i.e., by recognizing task-specific data with a recognizer developed for a different task. We chose to evaluate the performance of broadcast news acoustic and language models, on three commonly used tasks: small vocabulary recognition (TI-digits), goal-oriented spoken dialog (ATIS), and read and spontaneous text dictation (WSJ). The broadcast news 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.

The next section provides an overview of the LIMSI broadcast news transcription system used as our generic system. In Section 3, cross-task experiments serve to gain insight about the degree of genericity of the BN models. Then transparent adaptation techniques are shown to be effective in improving performance under cross-task conditions (Section 4). Finally multi-source training techniques are proposed and tested in Section 5 to improve the genericity of the BN models.

2. SYSTEM DESCRIPTION

The speech recognizer of LIMSI broadcast news transcription system [3] uses continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on large text corpora for language modeling. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained by means of a decision tree. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation,

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

Task	*BN ac. & lang. models*	*BN ac. & task lang. models*	*Task ac. & lang. models*
BN	13.6	13.6	13.6
TI-digits	17.5	1.7	0.4
ATIS	22.7	4.7	4.1
WSJ read	11.6	9.0	7.6
WSJ spon	12.1	13.6	15.3

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

3) final hypothesis generation. The initial hypotheses are used for cluster-based acoustic model adaptation using the MLLR technique [8] prior to word graph generation. A 3-gram LM is used in the first two decoding steps. The final hypotheses are generated with a 4-gram LM and acoustic models adapted with the hypotheses of step 2.

In the baseline system used in DARPA evaluation tests, the acoustic models were trained on about 150 hours of audio data from the DARPA/LDC Hub4 Broadcast News corpus [6]. Gender-dependent acoustic models were built using MAP adaptation of SI seed models for wideband and telephone band speech [4]. The models contain 28000 position-dependent, cross-word triphone models with 11700 tied states and approximately 360k Gaussians [5]. The baseline language models are obtained by interpolation of models trained on newspaper and newswire texts, commercial transcripts and transcriptions of acoustic training data. The recognition vocabulary contains 65120 words with on average 1.2 pronunciations per word. The pronunciations make use of a set of 48 phones, where 3 units represent silence, filler words, and breath noises.

The LIMSI 10xRT system had a word error of 17.1% on the 1999 NIST evaluation set and can transcribe unrestricted broadcast data with a word error of about 20% [5].

3. ASSESSING GENERICITY

Our first step in developing a generic speech transcription engine is to assess the system under cross-task conditions, i.e., by recognizing task-specific data with a recognizer developed for a different task. For the small vocabulary recognition task, experiments are carried out on the adult speaker portion of the TI-digits corpus [9] (17k utterances from 225 speakers). The vocabulary contains the digits '1' to '9', plus 'zero' and 'oh'. The database contains about 7 hours of speech, equally divided between training and test. 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%. Our task-specific recognition system has only 108 context-dependent phone models due to the low phonemic coverage of the digits. The task-specific LM is a simple grammar allowing any sequence of up to 7 digits. Our task-dependent system WER is 0.4%.

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 [2] is used for testing purposes. The test data amounts to nearly 5 hours of speech from 24 speakers recorded with a close-talking microphone. Around 40h of speech data are available for training. The 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 trigram back-off language model was estimated on the transcriptions of the training utterances. The lexicon contains 1300 words, with compounds words for multi-word entities in the air-travel database (city and airport names, services etc.). The WER of our task-dependent system is 4.1%.

For the dictation task, the *Wall Street Journal* continuous speech recognition corpus [12] is used, abiding by the ARPA 1995 Hub3 test conditions. The acoustic training data consist of 100 hours of speech from a total of 355 speakers taken from the WSJ0 and WSJ1 corpora. The Hub3 baseline test data consists 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% [14]. A contrastive experiment is carried out with the WSJ93 Spoke 9 data comprised of 200 spontaneous sentences spoken by journalists [7]. The best performance reported in the 1993 evaluation on the spontaneous data was 19.1% [13], however lower word error rates have since been reported on comparable test sets (14.1% on the WSJ94 Spoke 9 test data). 21k 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 trigram back-off model obtained by interpolating models trained on different data sets (training utterance transcriptions and newspapers texts). 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 reference BN transcription task, we follow the conditions of the 1998 ARPA Hub4E evaluation [10]. The acoustic training data is comprised of 150 hours of North American TV and radio shows. The LIMSI BN system obtained a 13.6% WER in the 1998 evaluation.

Three sets of experiments are reported in Table 1. 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. 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 1, right) and mixed (Table 1, middle) conditions, that the BN acoustic models are relatively generic. 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 task-

Task	BN acoust. models	BN acoust., unsupervised adaptation		BN acoust., supervised adaptation	
		MAP	MLLR+MAP	MAP	MLLR+MAP
TI-digits	1.7	0.8	0.8	0.5	0.5
ATIS	4.7	4.8	4.5	3.2	3.2
WSJ read	9.0	7.3	6.9	6.7	6.5
WSJ spon	13.6	12.6	11.9	11.6	11.0

Table 2: Word error rates (%) for TI-digits, ATIS, WSJ read and WSJ spontaneous 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. Two different adaptation schemes have been evaluated: MAP alone or MLLR followed by MAP.

specific LMs are more closely matched to BN and only a relative WER reduction of 20% is obtained. On the spontaneous journalist dictation test data there is even an increase in WER using the WSJ LMs, which can be attributed to a better modelization of spontaneous speech effects (such as breath and filler words) in the BN models.

4. TASK ADAPTATION

The experiments reported in the previous section show that while recognition with the reference BN acoustic models gives relatively competitive results, the WER on the targeted tasks can still be improved. Since the goal is to minimize the cost and effort in tuning to a target task, we are investigating methods to transparently adapt the reference acoustic models. Transparent means that the procedure is automatic and can be carried out without any human expertise. The approach proposed earlier is applied, i.e., the reference BN system is used to transcribe the training data of the target task. This supposes of course that audio data have been collected. However, the data collection cost 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 three tasks. The 100 hours of the WSJ data were transcribed using the BN acoustic and language models. Due to time constraints, only 26 of the 40 hours of the ATIS training data were transcribed. For TI-digits, the training data was transcribed using a mixed configuration, combining the BN acoustic models with the simple digit loop grammar as in this case using a task-specific LM is costless.

In order to assess the quality of the automatic transcription, the system hypotheses were scored against the manually provided training transcriptions. The resulting word error rates on the training data are 11.8% for WSJ, 29.1% for ATIS and 1.2% for TI-digits. For completeness, the task-specific audio data and associated transcriptions were used to carry out supervised adaptation of the BN models.

Task	BN ac. & lang models	BN unsupervised adapt ac. & lang models
ATIS	22.7	15.8
WSJ read	11.6	7.8
WSJ spon	12.1	11.4

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

Both the MAP and MLLR adaptation techniques were applied. MLLR was used with a global transformation followed by phone-based transformations. Combining the two adaptation techniques was also investigate by first adapting the BN models using MLLR, followed by MAP adaptation. Recognition tests were carried out under mixed conditions, using the adapted acoustic models and the task-dependent LM. The word error rates obtained with the task-adapted BN models are given in Table 2 for the four test sets. Using acoustic models with unsupervised adaptation, the performance is improved in every case: TI-digits (53% relative), ATIS (4% relative), WSJ (23% relative) and spontaneous WSJ (11% relative). As expected, the results using the manual transcriptions of the adaptation data from the targeted tasks to carry out supervised model adaptation are substantially better than unsupervised adaptation for both the TI-digits (37% relative) and ATIS (29% relative) tasks. Smaller relative gains of about 5% are obtained for the spontaneous dictation task and for the read WSJ data. The gain appears to be correlated with the WER of the transcribed data: the difference between BN and task-specific models is smaller for WSJ than for ATIS and TI-digits. Supervised adaptation using the MLLR technique, followed MAP adaptation is seen to yield equivalent or better results than MAP alone.

The performance improvement due to acoustic model adaptation is also seen when BN language models are used. Table 3 reports results using the BN language models instead of task-specific ones. For WSJ read and spontaneous, the WERs (7.8% and 11.4%) are even lower than for the task-specific system (see Table 1). The linguistic proximity of the BN and WSJ tasks can largely explain these results. Even for the ATIS task, a 30% relative WER reduction is observed. The TI-digits task is the only task for which

Task	*Task-id acoust model*		*Task specific acoust models*
	Pooling	*Sequential*	
BN (10×RT)	14.5	14.8	14.2
TI-digits	0.7	0.6	0.4
ATIS	3.1	3.6	4.1
WSJ read	6.7	7.4	7.6

Table 4: Word error rates (%) for BN98, TI-digits, ATIS and WSJ read95 test sets after recognition with three different configurations, all using task-specific lexica and LMs and MAP adapted BN acoustic models: (left) pooled data acoustic model adaptation, (middle) sequential acoustic model adaptation and (right) supervised task-dependent adaptation.

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

5. IMPROVING GENERICITY

In the previous section, the quality of the acoustic models was improved by confronting them with task-specific audio data. In this section methods to improve genericity of the models via multi-source training are investigated. This can be done in a variety of ways – by pooling data, by interpolating models or via single or multi-step model adaptation. Our aim is to obtain generic models results which are comparable to the respective task-dependent results for all tasks under consideration.

The most straightforward approach consists of merging the data from all of the tasks. The data pool is then used to adapt the BN acoustic models (the BN data are not included). Instead of pooling the data, a multi-step method can be considered, where the BN models are sequentially adapted with data from the other tasks. In this work, the BN acoustic models are first adapted with the WSJ data, then with ATIS data and finally with TI-digits data. In these experiments MAP adaptation was used.

The results for both training schemes are given in Table 4 using task-specific lexica and LMs, and supervised acoustic model adaptations. For the sequential adaptation, results are reported with the final model set. Compared to the results obtained with task-dependent acoustic models (Section 4), both the data pooling and sequential adaptation schemes lead to better performance for ATIS and WSJ read, with slight degradations for BN and TI-digits.

6. CONCLUSIONS

In this paper, new insights have been gained on 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. Models from the broadcast news task were chosen as reference models since they cover 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.

It has been demonstrated 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.

Finally, attempts have been made to enhance the genericity of the acoustic models. Multi-source training has been shown to improve the accuracy of the generic models, yielding recognition performance comparable or better than that obtained with task-specific models.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

[13] 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*, 81-88, 1994.

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