TOWARDS TASK-INDEPENDENT SPEECH RECOGNITION*

Fabrice Lefevre, Jean-Luc Gauvain and Lori Lamel

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{lefevre,gauvain,lamel}@limsi.fr

ABSTRACT

Despite the considerable progress made in the last decade, speech recognition is far from a solved problem. For instance, porting a recognition system to a new task (or language) still requires substantial investment of time and money, as well as expertise in speech recognition. This paper takes a first step at evaluating to what extent a generic state-of-the-art speech recognizer can reduce the manual effort required for system development.

We demonstrate the genericity of wide domain models, such as broadcast news acoustic and language models, and techniques to achieve a higher degree of genericity, such as transparent methods to adapt such models to a specific task. This work targets three tasks using commonly available corpora: small vocabulary recognition (TI-digits), text dictation (WSJ), and goal-oriented spoken dialog (ATIS).

1. INTRODUCTION

The last decade has witnessed major advances in the capability and performance of the speech recognition systems, derived from the improved accuracy and complexity of the statistical models. The availability of large spoken and text corpora for training, and the wide availability of more efficient and cheaper computational means have enabled the development and implementation of better training and decoding algorithms. Most notably there has been a move from recognition of speaker specific data produced with the purpose of being recognized to so-called "found data" such as radio and television broadcasts.

However, there are many outstanding research issues that need to be addressed before the speech recognition problem can be considered as "solved". Recognizer performance is still very sensitive to the environmental conditions and speaking style: channel type and quality, speaker characteristics, and background noise have an large 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. Most recognition systems are finely tuned so as to achieve reasonable performance for a particular application. For instance, a system trained on a read-speech corpus is unlikely to provide near optimal performance on a medium-size dialog task such as ATIS. Therefore the commonly adopted approach is to develop a system for each specific task. Given the very large number of potentially different situations in which speech recognizers can be used, this lack of genericity acts inhibits the widespread use of speech recognition technology.

The overall objective of the work presented here is the development of "generic" core speech recognition technology. 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, the genericity of the models is evaluated. From this perspective, a first experiment consists of assessing the recognition performance obtained under cross-task conditions, i.e., by recognizing task-specific data with a recognizer developed for a different task. In choosing the models, several considerations are important. The task should be somewhat general, covering a wide variety of linguistic and acoustic events in the language, so as to ensure reasonable coverage of the target task. There should be sufficient acoustic and linguistic training data available so that the models are accurate and cover a wide range of speaker and language characteristics.

From these considerations, we selected the LIMSI broadcast news (BN) transcription system to use as a 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.

After the performance of the BN system on the targeted tasks is evaluated, an attempt is made to improve its generic-

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

ity through transparent methods. Recent studies [5, 7] have proposed solutions for reducing the development cost for the BN task. In this work we attempt to apply the same approach using available task-specific training data in an unsupervised manner. This cross-task unsupervised acoustic adaptation uses only the raw audio training data for the target tasks, i.e., the manual orthographic transcriptions provided with the corpora are ignored. Since no manual transcription is required, this approach is much less costly than traditional task-specific training.

In the next section, the different tasks and their associated corpora are presented, followed by a description of the reference broadcast news transcription system. Experimental results are given for cross-task recognition in Section 4 and after unsupervised acoustic model adaptation in Section 5.

2. TASK DESCRIPTIONS

Two main criteria have guided the choice of the tasks studied: they should be realistic enough and task-specific data should be available. Four widely used tasks have been retained for this study: small vocabulary recognition (TI-digits), dictation of texts (WSJ), goal-oriented human-machine spoken dialog (ATIS) and broadcast news transcription (Hub4E). Although also widely used in the past, we consider the Resource Management and TIMIT corpora to be less realistic and therefore have not used them in this investigation. The characteristics of the tasks and 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 [9], 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%.

For the dictation task, the *Wall Street Journal* continuous speech recognition task [11] is used, along with 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. A contrastive experiment is carried out with the WSJ93 Spoke 9 data comprised of 200 spontaneous sentences spoken by journalists [6]. The best performance reported in the 1993 evaluation on the spontaneous data was 19.1% [12], however lower word error rates have since been reported on comparable test sets (14.1% on the WSJ94 Spoke 9 test data).

Finally, 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*) [1] used for testing purposes. This data contains 981 utterances (nearly 5 hours of speech) from 24 speakers recorded with a close-talking microphone. The training corpus contains about 100h of speech data. The word error rates for this task in the 1994 evaluation were mostly in the range of 2.5% to 5%, which we take as state-of-the-art for this task.

The reference task is BN transcription, with the conditions are those used in the 1998 ARPA Hub4E evaluation (*BN98*) [10]. The acoustic training data is composed of 150 hours of North-American TV and radio shows. The test data are annotated accordingly to six focus-conditions corresponding to particular combinations of acoustic attributes and speaking styles. The intrinsic variety of conditions present in the BN data makes it a rather logical choice for the reference system. The best overall result on the 1998 baseline test was 13.5%.

3. BN REFERENCE SYSTEM

The LIMSI broadcast news transcription system has two main components, an audio partitioner and a word recognizer. Data partitioning [2] serves to divide the continuous audio stream into homogenous segments, associating appropriate labels for cluster, gender and bandwidth with the segments. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. The initial hypotheses are used for cluster-based acoustic model adaptation using the MLLR technique [8] prior to word graph generation. A 3-gram LM is used in the first two decoding steps. The final hypotheses are generated with a 4-gram LM and acoustic models adapted with the hypotheses of step 2.

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. The same basic acoustic model training procedure is used to build the task-dependent acoustic models.

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) [4]. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second derivatives. Gender-dependent acoustic models are estimated using MAP adaptation of SI seed models for wideband and telephone band speech [3].

For computational reasons, a small set of acoustic models is used in the first step to generate the initial hypotheses. These position-dependent, cross-word triphone models cover 5650 contexts, with 6275 tied states and 16 Gaussians per state. For trigram decoding a larger set of 28064

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
WSJ	95	News Dictation	100h (355)	45mn (20)	Newspaper, newswire	6.7
S9_WSJ	93	Spontaneous Dictation	" "	43mn (10)	" "	19.1
ATIS	93	H-M Dialog	100h (137)	5h (24)	Transcriptions	2.5

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

	BN models					BN Ac. models + Task LMs					Task-dependent models				
Test Set	*Corr*	*Sub*	*Del*	*Ins*	*Err*	*Corr*	*Sub*	*Del*	*Ins*	*Err*	*Corr*	*Sub*	*Del*	*Ins*	*Err*
BN98	88.3	8.8	2.9	1.8	13.6	88.3	8.8	2.9	1.8	13.6	88.3	8.8	2.9	1.8	13.6
TI-digits	86.3	11.9	1.8	3.8	17.5	98.5	0.7	0.8	0.3	1.7	99.8	0.1	0.2	0.2	0.4
WSJ95	90.9	8.1	1.0	2.5	11.6	92.5	6.6	0.9	1.5	9.0	93.3	5.7	1.0	1.4	8.2
S9_WSJ93	91.7	6.6	1.7	3.7	12.1	91.2	7.0	1.8	4.8	13.6	91.0	7.3	1.7	6.3	15.3
ATIS94	81.5	17.0	1.5	4.3	22.7	96.8	2.5	0.7	1.5	4.7	97.4	2.1	0.5	1.8	4.4

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

position-dependent, cross-word triphone models with 11700 tied states was used. Finally, the 4-gram decoding uses a 32 Gaussians per state version of the larger model set.

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 recognition vocabulary contains 65,120 words. A pronunciation graph is associated with each word, represented using a set of 48 phones (specific phone symbols are used to explicitly model filler words and breath noises).

4. CROSS-TASK RECOGNITION

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. Finally, in the third experiment set task-dependent models were developed and evaluated for each task.

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 task-specific LM for the TI-digits is a simple grammar allowing any sequence of up to 7 digits.

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 that the BN acoustic models are relatively generic by comparing the task-dependent (Table 2, middle) and mixed conditions (Table 2, right). These models seem to be a good start towards truly task-independent models.

The word error increase observed in cross-task conditions (Table 2, left) is large for the TI-digits and ATIS tasks. For both of these tasks, the mixed model experiment shows that the gap in performance is mainly due to the language models. For the WSJ95 data the gain obtained by using task-specific training data is smaller than for the other tasks, due to the similarity with the BN task. The results on the S9_WSJ93 spoke (spontaneous journalist dictation) show 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.

5. UNSUPERVISED ADAPTATION

The experiments reported in the previous section show that while direct recognition with the reference BN acoustic models gives relatively competitive results, the WER on the targeted tasks can be improved. As we are targeting no or minimal task-dependent tuning, we investigated a transparent method based on unsupervised adaptation of the reference acoustic models.

The basic idea is to use the BN system to transcribe the task-specific training data of the destination task. Then, the acoustic models are adapted by means of a conventional

Test Set	*Adaptation*	*Corr*	*Sub*	*Del*	*Ins*	*Err*
WSJ95	MAP	92.0	6.9	1.1	2.1	10.0
WSJ95	MLLR	91.7	7.2	1.1	2.1	10.3
TI-digits	MLLR	99.0	0.4	0.6	0.3	1.3

Table 3: Word error rates (%) for the WSJ and TI-digits test data after unsupervised adaptation of the BN acoustic models using the respective training corpora.

adaptation technique such as MAP or MLLR. This supposes of course that audio data have been collected for the targeted task. However, the cost of collecting task training data is greatly reduced since no manual transcriptions are needed. We have chosen to adapt the original BN models with the task-specific data, thus ensuring the modelization of many different phonemic contexts.

The cross-task unsupervised adaptation is evaluated for both the TI-digits and WSJ tasks. About 25 hours of WSJ speech from 80 speakers was transcribed using the BN acoustic and language models. The WER measured on this data is 11.8% (using the manual transcripts for the reference). For TI-digits task, the training data was transcribed using a mixed configuration, combining the BN acoustic models with the simple digit loop grammar. A WER of 1.2% was measured on the TI-digits training data.

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, the speaker ids have been 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 is automatically classified by decoding each utterance twice, once with each set of gender-dependent models. Then, the utterance gender is determined based on the best global score between the male and female models (99.0% correct classification).

The test set WERs obtained with the task-adapted BN models are given in Table 3. The MLLR technique has been used for the TI-digits task and the test was carried out under mixed conditions (i.e., with the task-dependent LM). It can be observed that gains are obtained in all cases (24% relative for TI-digits and 14% relative for WSJ95). On the WSJ data, MAP obtains a larger improvement in WER than MLLR (11% relative). This may be because only one global transformation was used for the MLLR adaptation.

6. SUMMARY

This paper has explored the genericity of state-of-the-art speech recognition systems, by testing a relatively broad system on data from three tasks ranging in complexity. The initial set of generic models were taken from the broadcast news task, since this task covers a wide range of acoustic and linguistic conditions. The acoustic models are shown to be relatively task-independent as only a small increase in word error is obtained under cross task conditions when using task-dependent language models. There remains a large difference in performance on the digit recognition task which can be attributed to the limited phonetic coverage of this task, as well as to the particular evaluation corpus with ensured complete adherence with the task. On a spontaneous WSJ dictation task, the BN models are more robust to deviations in speaking style than the read-speech WSJ models. Comparing performance with and without a task-specific language model, shows that this is due both to the acoustic and language models.

We also have shown that unsupervised acoustic model adaptation can reduce the performance gap between task-independent and task-dependent acoustic models. For the WSJ95 task the relative error increase is reduced from 22% to 8%. Unsupervised adaptation is less effective for the digits task, confirming our earlier observation that digits are special case.

REFERENCES

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

[2] 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.

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

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

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

[6] 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.

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

[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), pp. 171-185, 1995.

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

[10] 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.

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

[12] 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.