Advances in Transcription of Broadcast News and Conversational Telephone Speech Within the Combined EARS BBN/LIMSI System

Article *in* IEEE Transactions on Audio Speech and Language Processing · October 2006

DOI: 10.1109/TASL.2006.878257 · Source: IEEE Xplore

CITATIONS
72

READS
308

15 authors, including:

Jean-Luc Gauvain
Computer Science Laboratory for Mechanics and Engineering Sciences
316 PUBLICATIONS **12,979** CITATIONS
SEE PROFILE

Thomas Colthurst
Google Inc.
25 PUBLICATIONS **744** CITATIONS
SEE PROFILE

O. Kimball
Raytheon Technologies
60 PUBLICATIONS **2,422** CITATIONS
SEE PROFILE

Lori Lamel
French National Centre for Scientific Research
423 PUBLICATIONS **14,669** CITATIONS
SEE PROFILE

Advances in Transcription of Broadcast News and Conversational Telephone Speech Within the Combined EARS BBN/LIMSI System

Spyros Matsoukas, *Member, IEEE*, Jean-Luc Gauvain, *Member, IEEE*, Gilles Adda, Thomas Colthurst, Chia-Lin Kao, *Member, IEEE*, Owen Kimball, *Member, IEEE*, Lori Lamel, *Member, IEEE*, Fabrice Lefevre, Jeff Z. Ma, *Member, IEEE*, John Makhoul, *Fellow, IEEE*, Long Nguyen, *Member, IEEE*, Rohit Prasad, *Member, IEEE*, Richard Schwartz, Holger Schwenk, *Member, IEEE*, and Bing Xiang, *Member, IEEE*

***Abstract*—This paper describes the progress made in the transcription of broadcast news (BN) and conversational telephone speech (CTS) within the combined BBN/LIMSI system from May 2002 to September 2004. During that period, BBN and LIMSI collaborated in an effort to produce significant reductions in the word error rate (WER), as directed by the aggressive goals of the Effective, Affordable, Reusable, Speech-to-text [Defense Advanced Research Projects Agency (DARPA) EARS] program. The paper focuses on general modeling techniques that led to recognition accuracy improvements, as well as engineering approaches that enabled efficient use of large amounts of training data and fast decoding architectures. Special attention is given on efforts to integrate components of the BBN and LIMSI systems, discussing the tradeoff between speed and accuracy for various system combination strategies. Results on the EARS progress test sets show that the combined BBN/LIMSI system achieved relative reductions of 47% and 51% on the BN and CTS domains, respectively.**

***Index Terms*—Hidden Markov models (HMMs), large training corpora, speech recognition, system combination.**

I. Introduction

In May 2002, DARPA initiated a five-year research program called EARS (Effective, Affordable, Reusable, Speech-to-text). The major goal of the program was to reduce recognition word error rates (WERs) for broadcast news (BN) and conversational telephone speech (CTS) by a factor of five in five years to reach the 5%–10% range while running in real-time on a commodity computer with only a single processor. The drive to lower WER and to real time was in several phases with milestones to be achieved at the end of each phase. For example, the performance target for BN systems developed during the second phase of the program (2003–2004) was a WER of 10% at a speed of ten times real-time (10 $\times$ RT). Progress was measured on a "Progress Test" in English which remained fixed for the five-year duration of the program. In addition, there were "Current Tests" in each of the three languages (English, Arabic, and Mandarin), which changed every year. These yearly evaluations are referred to as the Rich Transcription benchmark tests (RT02, RT03 and RT04). Collaboration across sites was strongly encouraged. BBN and LIMSI have been working closely together and, wherever possible, submitted joint results.

For the BN domain, the transcription systems must be able to deal with the nonhomogeneous data found in broadcast audio, such as a wide variety of speakers and speaking styles, changing speakers, accents, background conditions, and topics. The challenges for CTS at the acoustic level concern speaker normalization, the need to cope with channel variability, spontaneous speech, and the need for efficient speaker adaptation techniques. On the linguistic side, the primary challenge is to cope with the limited amount of language model training data. Although substantially more CTS training data was made available under the EARS program (see Section II), appropriate textual data for training the language models are difficult to obtain.

There are notable differences in speaking style observed in CTS and BN. Broadcast speech is much closer to written language than conversational speech is, where different social conventions are observed. For CTS, the acoustic conditions are quite varied. The speech quality is affected by a variety of different types of telephone handset, the background noise (other conversations, music, street noise, etc.), as well as a much higher proportion of interruptions, overlapping speech, and third person interjections or side conversations. In terms of linguistic content, there are many more speech fragments, hesitations, restarts and repairs, as well as back-channel confirmations to let each interlocutor know the other person is listening.

The first-person singular form is much more predominant in conversational speech. Another major difference from BN is that some interjections such as "uh-huh" and "mhm" (meaning yes) and "uh-uh" (meaning no) that are considered as nonlexical items in BN, need to be recognized since they provide feedback in conversations and help maintain contact. The word "uhhuh," which serves both to signal agreement and a back-channel "I'm listening," accounts for about 1% of the running words in the CTS data. The most common word in the English CTS data, "I," accounts for almost 4% of all word occurrences, but only about 1% of the word occurrences in BN.

Manuscript received October 15, 2005; revised April 10, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mary P. Harper.

S. Matsoukas, T. Colthurst, C.-L. Kao., O. Kimball, J. Ma, J. Makhoul, L. Nguyen, R. Prasad, R. Schwartz, and B. Xiang are with BBN Technologies, Cambridge, MA 02138 USA (e-mail: smatsouk@bbn.com).

J.-L. Gauvain, G. Adda, L. Lamel, F. Lefevre, and H. Schwenk are with the Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI), National Center for Scientific Research, 75794 Paris Cedex 16, France.

Digital Object Identifier 10.1109/TASL.2006.878257

This paper reports on the development and evaluation of the combined BBN/LIMSI English systems for BN and CTS. Most of the development work was focused on the English language, in part due to the larger amount of audio and textual data available for this language, and in part because the official progress test was on English data. However, our experience thus far indicates that at today's word error rates, the techniques used in one language can be successfully ported to other languages, and most of the language specificities concern lexical and pronunciation modeling.

The EARS program led to significant advances in the state-of-the-art for both BN and CTS transcription, including the development of new training methods to deal with the very large quantities of audio data, to semi- or lightly supervised training methods to reduce the need for manual annotations, and to innovative approaches for system combination (developing complementary systems, cross system adaptation, cascade and parallel architectures for combination) to meet the time constraint requirements.

The remainder of this paper is as follows. In the next section the training and test corpora used in the experiments are described. Sections III and IV highlight some of the main features, development work, and specificities for the BBN and LIMSI component systems, and Section V discusses other ideas that were tried but did not become part of the final evaluation systems. Section VI describes the different strategies that were explored for system combination, and Section VII gives joint BBN/LIMSI results on the progress and evaluation test sets.

II. Training and Test Corpora

In the beginning of the EARS program, the available acoustic training corpora consisted of approximately 230 h of CTS data (Switchboard-I plus Callhome conversations), and 140 h of BN data, both carefully transcribed. Since then, the CTS acoustic training corpus has grown to approximately 2300 h of speech with incremental additions of Switchboard-II and Fisher conversations [1]. It is worth noting that the reference transcripts for almost all of the additional material was obtained via quick transcription [2], with time segmentation provided automatically by the BBN system. The BN acoustic training corpus enjoyed a similar increase in size; however, in that case the additional material was obtained via light supervision methods [3], [4] from a large pool of closed-captioned TV shows (approximately 9000 h).

Both BBN and LIMSI systems also made use of large amounts of text data for language model (LM) training. The CTS LM training included 530 M words of web data released by the University of Washington (UW), Seattle, [5], 141 M words from BN data, 47 M words of archived text from CNN and PBS, and 2 M words from the closed captions of the Topic Detection and Tracking 4 (TDT4) database. The BN LM training consisted of approximately 1 billion words of text, including the American English GigaWord News corpus, commercial transcripts from PSMedia, and CNN web archived transcripts.[1]

Several test sets were used to evaluate system performance during the EARS program. Although overall system performance was benchmarked yearly by NIST on the designated progress test sets, day to day research was evaluated on certain development sets. In addition, less frequent tests were performed on "evaluation" sets, in order to validate gains from various modeling approaches. Tables I and II list the characteristics of the test sets that were used for the experimental results reported in this paper.

[1]Most of the data are distributed by the Linguistic Data Consortium.

TABLE I
CTS Test Sets

Test set	Type	#hours	#words
Eval01	Development	6	63k
Eval02	Validation	6	64k
Eval03	Development	6	76k
Dev04	Development	3	38k
Eval04	Validation	3	37k

TABLE II
BN Test Sets

Test set	Type	#hours	#words
Dev03	Development	3	25k
Dev04	Development	3	25k
Eval04	Validation	6	46k

III. BBN System Highlights

A. Decoding Architecture

The BBN system uses a multipass decoding strategy in which models of increasing complexity are used in successive passes in order to refine the recognition hypotheses [6]. The first pass is a fast-match search [7] performed in the forward direction, using a bigram language model and a composite within-word triphone hidden Markov model (HMM). Tying of the HMM states is performed via a linguistically guided decision tree for each phoneme and state position. In one configuration, all triphones of a given phoneme share the same set of Gaussian components, while in another, the sharing of Gaussians is done for each phoneme and state position. In both cases, the mixture weights are shared based on the decision tree clustering. We use the terms "phonetically tied mixture" (PTM) and "state tied mixture" (STM) to refer to these two types of models, respectively. The output of the forward pass consists of the most likely word ends per frame along with their partial forward likelihood scores. This set of choices is used in a subsequent backward pass to restrict the search space, allowing for less expensive decoding with more detailed acoustic and language models.

The backward pass is a time-synchronous beam search, employing an approximate trigram language model and within-word quinphone State clustered tied mixture (SCTM) HMMs. State tying in the SCTM model is determined based on a linguistically guided decision tree, similar to the PTM/STM. In the SCTM case though, the decision tree is grown in two steps. In the first step, a high threshold on the state cluster occupancy counts is set, and the resulting state clusters determine the sharing of the Gaussian components (codebooks). In the second step, each codebook cluster is divided further by the decision tree, using a lower occupancy threshold, to determine the sharing of the mixture weights. The output of the backward pass is either a hypothesis N-best list, or a word lattice.

The decoding is completed with a rescoring pass, operating on the N-best/lattice. This makes use of between-word quinphone SCTM acoustic models and more accurate language models (e.g., fourgrams or part-of-speech smoothed trigrams).

It should be noted that each of the above three passes can read in acoustic feature and/or model transformations that are performed on the fly, on a per speaker basis, for the purpose of speaker adaptation. Thus, a full decoding experiment typically runs in an iterative manner, interleaving speaker adaptation with recognition in order to provide the best output.

The decoding process, when operating as a single system, is repeated three times. First, speaker-independent (and gender-independent) acoustic models are used in the decoding to generate hypotheses for unsupervised adaptation. Then, the decoding is repeated but with speaker-adaptively trained acoustic models that have been adapted to the hypotheses generated in the first stage. The last decoding is similar to the second, but acoustic models are adapted to the second stage's hypotheses using a larger number of regression classes.

B. RT02 Baseline Systems

The BBN RT02 baseline CTS system [8] used vocal tract length normalization (VTLN) [9] operating on LPC-smoothed spectrum, producing 14 cepstral coefficients plus normalized energy per frame of speech (25-ms window, 10-ms frame step). Mean and covariance normalization were applied to the cepstra on a conversation side basis, to reduce variability due to the channel/speaker. The base frame features were augmented with their first, second, and third time derivatives to produce a 60-dimensional feature vector that was then projected to 46 dimensions using linear discriminant analysis (LDA). A global maximum-likelihood linear transform (MLLT) [10] was applied to the resulting features in order to make them more suitable for modeling with diagonal covariance Gaussian distributions. Gender-dependent (GD) between word quinphone SCTM models were estimated from the acoustic training data using maximum likelihood (ML), both with and without speaker adaptive training (SAT) [11]. A special form of SAT was employed, in which the matrix applied to the Gaussian mean vectors was diagonal. This had been shown previously to work as well as using a full matrix. In addition to the ML models, a set of maximum mutual information (MMI) models was estimated, using N-best lists to represent the set of confusable hypotheses. The average number of Gaussians in a GD SCTM model was about 422 k. Compound words were used in both acoustic and language model training. A trigram LM of about 17 M trigrams was used in recognition. In addition, a part-of-speech smoothed trigram LM was used in N-best rescoring.

The BBN RT02 baseline BN system was based on the 1999 $10 \times$ RT Hub-4 BN system [12] with small changes in the automatic segmentation procedure. GD, band-dependent (BD), between-word SCTM models were estimated with ML (featuring approximately 256 k Gaussians per model). No SAT was used. A small set of compound words were included in the acoustic and language model training. A trigram LM with 43 M trigrams was used for both N-best generation and N-best rescoring. To fulfill the run-time requirement of $10 \times$ RT, the BBN system used shortlists for reducing Gaussian computation, took advantage of memory caching during the Gaussian computation in the forward pass, and spread the grammar probability to each phoneme of a word to allow effective pruning with tighter beams [13].

Recognition results using both BBN RT02 systems were selected as the baselines to measure progress in reduction of the WER for the EARS program. For the BN task, the baseline WER, measured on the *Progress Test*, was 18.0%. For the CTS task, instead of using the manual segmentation of the test material that had been the norm for all pre-EARS CTS research programs, we used an automatic segmentation generated by an algorithm developed at MIT Lincoln Laboratory that was available at that timeframe. The CTS baseline WER was 27.8%.

C. System Improvements

With the beginning of the EARS program, a number of improvements were incorporated to the BBN system. In this section, we highlight the most notable improvements, and give results that show their effect on recognition accuracy.

1) Automatic Segmentation: One of the first priorities for the BBN/LIMSI team in the EARS program was the implementation of a robust automatic audio segmenter for CTS. Recall that CTS recordings are done in stereo, with each conversation side stored in a separate channel. One could try to segment the two sides independent of each other. However, this approach often suffers from poor segmentation performance in regions of crosstalk due to the leakage of speech from one channel into the other during the recording. To avoid this problem, BBN developed a segmentation procedure that processes both sides of the conversation simultaneously [14]. The algorithm uses an ergodic HMM with four states, corresponding to the four combinations of speech or nonspeech events on each side of the conversation. The observations of the HMM consist of joint cepstral features from both conversation channels. Experiments on the CTS Eval02 test set show that this segmentation procedure degrades WER only by 0.2%–0.4% absolute compared to using the manual segmentation.

2) Speaker Adaptive Training: Two SAT improvements were developed during the EARS program. The first one was to use constrained maximum likelihood linear regression (CMLLR) [15] adaptation in training, with one transformation per speaker. The computational advantage of this method, as described in [15], is that one can apply the speaker transforms to the observations and then build a regular system in the transformed feature space. Both ML and MMI estimation can be used to train the final models. CMLLR-SAT was found to be better than SAT with diagonal transforms on CTS data, and it also provided a gain in the BN system. The second SAT improvement, called heteroscedastic linear discriminant analysis (HLDA)-SAT [16], was motivated from initial experiments that showed an improvement in the HLDA objective function, when the input feature space was normalized to some extend in order to reduce interspeaker variability (e.g., through the use of VTLN). In HLDA-SAT, the base cepstra and energy are transformed for each training speaker in order to maximize the likelihood of the data with respect to a canonical model. This model has the same structure as the one that is typically used in HLDA estimation, i.e., it consists of a single full covariance Gaussian for each codebook cluster. Starting with

TABLE III
SAT RESULTS ON BN DEV03 TEST SET, USING ML MODELS TRAINED ON 140 h (3-GRAM LM)

Model	Test Adaptation	WER
SI	no	18.6
SI	yes	15.7
CMLLR-SAT	yes	15.2
HLDA-SAT	yes	14.6

speaker-dependent transforms set to the identity matrix, the procedure runs a few iterations of interleaved transform and model updates. The resulting canonical model is then used in the transformed space in order to estimate the global HLDA projection. After that, a typical CMLLR-SAT is carried away.

HLDA-SAT was shown to provide a significant gain in the BN system, on top of the gain from CMLLR-SAT, as shown in Table III. HLDA-SAT was also applied to CTS data, but with no significant gain over the CMLLR-SAT CTS baseline. We believe that this is due to the fact that the CTS baseline already applied speaker/channel normalization techniques (VTLN, covariance normalization) on the observations prior to the estimation of the global HLDA projection.

It is worth mentioning that the use of HLDA-SAT eliminated the need to train gender-dependent and band-dependent acoustic models. A single HLDA-SAT acoustic model was found to be as good as GD, BD HLDA-SAT models in BN experiments.

3) Increasing LM Size: Early in the development of the EARS BBN system, it was found that there was a gain in recognition accuracy from using unpruned trigram or fourgram LMs. Based on that result, a procedure was designed that allowed efficient training of a very large LM and fast usage of the stored LM statistics for exact N-best rescoring. The procedure enumerates the set of unique n-grams found in the N-best list, and consults the stored LM sufficient statistics in order to compute a small LM that covers all the observed n-grams. A 0.4%–0.5% absolute gain was measured on both CTS and BN data from this improvement.

4) Adding More Acoustic Training Data: One of the largest improvements during the development of the EARS system was due to the use of a large acoustic training corpus. In CTS, the quickly transcribed data was automatically aligned with the audio using a version of the BBN system, and the resulting segments were included in both acoustic and language model training. Addition of the new data was done in two increments. In the first phase, approximately 80 h of Switchboard-II data were included in both acoustic and language model training, improving the CTS ML baseline from 27.8% to 25.8% (as measured on the Switchboard-II and Cellular portion of the 2001 Hub-5 Evaluation test set). During the second phase of the EARS program, a substantial amount of Fisher data was collected and transcribed, resulting in about 1930 h of speech (after the BBN post-processing). Including the Fisher data to the BBN AM and LM training resulted in additional gains, as shown in Table IV. Note that the WER improved on both the Switchboard and Fisher portions of the Eval03 test set, even though only Fisher data were added in training.

In BN, the use of the additional closed-captioned audio required light supervision methods. The closed captions were first

TABLE IV
ADDING 1930 h OF FISHER DATA TO 370 h OF SWITCHBOARD ACOUSTIC TRAINING. RESULTS ON CTS EVAL03 TEST SET, WITH ADAPTATION

AM	LM	#Gaussians	%WER (Eval03)		
			Swbd	Fsh	All
Swbd	Swbd	442k	28.6	20.3	24.6
Swbd	Swbd+Fsh	442k	27.3	19.0	23.3
Swbd+Fsh	Swbd	843k	26.5	19.2	23.0
Swbd+Fsh	Swbd+Fsh	843k	24.9	17.9	21.5

TABLE V
INCREASING AMOUNT OF BN TRAINING THROUGH LIGHT SUPERVISION. RESULTS ON BN DEV03 TEST SET, WITH ADAPTED HLDA-SAT MODELS (4-GRAM LM)

Source	#hours	#Gaussians	WER
H4	140	148k	12.6
H4+TDT4	297	354k	12.0
H4+TDT2,3,4	843	741k	11.0
H4+TDT2,3,4+BN03	1700	821k	10.5

normalized and used to produce a biased LM. A decoding was performed on the audio data using this targeted LM, and the resulting 1-best hypotheses were aligned with the closed captions to identify regions of agreement. Portions of the alignment with three or more consecutive word matches were extracted, along with the corresponding audio, to produce the extra acoustic training material. The procedure is described in more detail in [4]. One could argue that the light supervision method should not provide any significant improvements, since it extracts segments where the recognition output matches the available closed captions and, therefore, there are no errors to fix. This argument would be valid if a generic LM was used in recognition. Using a language model biased to the closed captions supports a weak acoustic model, and makes possible the accurate recognition of difficult audio portions, where recognition errors would normally occur with a generic LM.

The effect of increasing the acoustic training corpus in this way is shown in Table V, where it can be seen that the best result is obtained with 1700 h of training data, extracted by running light supervision on closed-captioned audio from all three TDT databases, as well as from BN material collected in 2003. Additional results were obtained by increasing the amount of training up to about 3000 h. However, no significant WER reduction was observed from this extra data.

It is important to note that, even though no special modeling techniques were needed in order to take advantage of the additional acoustic training material (we simply increased the number of states and/or Gaussians per state in the HMM), significant speed enhancements were necessary in the acoustic model estimation. In particular, to minimize input/output operations from/to the network file server, an effort was made to reduce the size of the cepstra and probabilistic frame-to-state alignment (label) files so that these files can easily fit on the local disk of each compute node in the batch queue. Cepstra files were reduced in size by a factor of 4, via linear quantization techniques. Label files were pruned aggressively to achieve a three-fold reduction in size. In addition, a single set of label files were used for training all three models (STM, within-word SCTM, and between-word SCTM) needed in our multipass decoding strategy.

TABLE VI
EFFECT OF DISCRIMINATIVE TRAINING OF ACOUSTIC MODELS. ADAPTED DECODING RESULTS

Trn. Method	CTS Dev04	BN Dev04
ML	18.4	12.3
MMI	16.2	11.3
MPE	15.7	11.6

TABLE VII
ADAPTED RECOGNITION RESULTS ON THE CTS DEV04 TEST SET, FOR COMPARING MODELS TRAINED USING LONG SPAN FEATURES WITH MODELS TRAINED USING FEATURE DERIVATIVES

LDA Pre-transform	%WER
Derivatives	15.4
Concatenated Frames	14.9

These file size reductions allowed distribution of the training data to each compute node, thereby enabling fast parallel processing with minimal network access. Further training speedups were obtained by implementing certain approximations in the Gaussian splitting initialization algorithm, such as splitting a larger number of Gaussian mixture components in each EM iteration, and partitioning the samples associated with a given codebook HMM state in regions in order to perform splitting within a region. More details can be found in [17].

5) Discriminative Training: Recall that the BBN RT02 CTS system used MMI training [18], but based on N-best lists rather than word lattices. During the first year of the EARS program, BBN implemented lattice-based MMI training, and obtained significant WER reductions, both on CTS and BN data, as shown in Table VI. The CTS acoustic models used in Table VI were trained on 2300 h of speech, while the BN models used approximately 1700 h.

During the second phase of EARS, minimum phoneme error (MPE) [19] training was also implemented. A particular form of MPE training, in which the objective function is smoothed with an MMI prior [20] was found to give optimal results. MPE-MMI resulted in 0.5% absolute gain on the CTS data, but no gain on the BN corpus compared to MMI. The degradation from MPE on the BN data might be due to overfitting, as the BN system used more Gaussians per training hour of speech than the CTS system, and MPE is known to be more sensitive to overfitting than MMI estimation.

Both MMI and MPE training perform forward–backward training passes on word lattices annotated with unigram language model probabilities. Using a weak language model during the discriminative training process is important, as described in [18], because a stronger language model trained on the same acoustic training data would produce excessive bias toward lattice paths that correspond to the reference transcript. BBN investigated an alternative approach, termed "held-out MPE" training [17], in which an initial MMI acoustic model was first trained on a subset of the CTS training data (800 h), and the remaining (1500 h) was treated as held-out set. The held-out data were also excluded from the training of a trigram LM. Word lattices were generated on the held out set using the initial acoustic model and the trigram LM, and a conventional MPE training was carried out on the (trigram) lattices. The idea was to perform MPE training in a scenario that simulated the recognition process on unseen data. Although held-out MPE did not improve recognition accuracy over regular MPE training, it performed equally well while using significantly smaller acoustic models.

6) Long-Span Features: Another research direction toward maximizing the benefit from the large acoustic training corpus was the design and estimation of feature transforms that incorporate large acoustic context information. It is well known that humans rely heavily on long acoustic context in order to recognize speech. HMMs, on the other hand, process speech on a frame to frame basis, with each frame typically spanning a range of 70–90 ms (through the use of time derivatives on cepstra and energy terms). Instead of simply augmenting the base frame features with their time derivatives prior to LDA, BBN explored the use of frame concatenation. Under this scheme, the observation vector at frame position t is constructed by first concatenating the energy and cepstral coefficients from frames $t-N,\ldots,t-1,t,t+1,\ldots,t+N$, and then projecting the spliced feature vector to a lower dimensional space via LDA. Up to 30 frames of context were considered. Significant WER reductions were obtained by incorporating context from 15-frame concatenation, as seen in Table VII.

However, extending the context to longer frame spans using standard LDA + MLLT techniques was found problematic, due to the suboptimality of the LDA criterion. Thus, a better procedure was developed that employed discriminative estimation of the large projection matrix, using the MPE criterion. MPE-HLDA [21] proved to be more robust than LDA, resulting in small but consistent WER improvements of about 0.3%–0.4% absolute on the CTS data.

Long span features were also applied to the BN domain but provided much smaller benefit. As a result, they were not included in the final BBN RT04 BN evaluation system.

7) Miscellaneous Decoding Improvements: Several decoding optimizations were implemented during the EARS program to ensure the decoding time to be within the $10\times$ RT or $20\times$ RT limits. The typical PTM model was replaced in the fast-match pass by a more detailed STM model, which, not only resulted in faster recognition (through tighter pruning), but also helped reduce the WER in the final rescoring pass. In addition, an improved method for reducing the Gaussian computation was implemented, based on [22]. Speaker adaptation was sped up significantly by using an approximation to the ML criterion, in which all dimensions of the observation vectors are assumed to have equal variance. All these optimizations are described in more detail in [17].

IV. LIMSI SYSTEM HIGHLIGHTS

A. From BN to CTS and Back to BN

The first CTS system developed at LIMSI used the same basic components as the LIMSI BN system [23]. Given the level of development of our BN models, it was of interest to benchmark the system on conversational telephone data without any modifications. These first experiments were done using the NIST Hub5 2001 test set (Eval01). Using both the BN acoustic and language models results in a word error rate of 51%. Keeping the same word list (the out of vocabulary (OOV) rate of the Eval01 data

is under 1% with the BN wordlist) and retraining the language model on the transcriptions of 230 h of Switchboard (SWB) data reduces the word error rate to 47.0%. Using both acoustic and LM models trained on the SWB data reduces the word error rate to 36%. This initial experiment demonstrates that a large part of the mismatch between the BN system and the SWB data is linked to the acoustic models, and that simply training models on the SWB data with our BN recipes was not enough to achieve good performance on conversational data.

Over the years of the DARPA EARS program, the CTS system was continually improved by incorporating a number of important features, in order to deal with the specificities of conversational telephone speech. The first set of additional features were VTLN, multiple regression class MLLR adaptation, neural-network language model, and consensus decoding with pronunciation probabilities. VTLN, which had not helped in the LIMSI BN transcription system, quite significantly improved the performance of the CTS system. Given that there is only a single speaker in each conversation side, on average there is more data available for MLLR adaptation, which is, therefore, more efficient with multiple regression classes. While for the BN task it is relatively easy to find a variety of task related texts, for conversational speech, since the only available source is the transcripts of the audio data, the generalization offered by a continuous space neural network LM [24] is of particular interest. Due to the higher word error rates on CTS data, lattice rescoring with consensus decoding and pronunciation probabilities performed significantly better than standard MAP decoding. The combination of these additional features reduced the WER to about 26% on the Eval01 test data (RT02 system).

Further improvements were achieved by incorporating gender-dependent VTLN, discriminative training (MMI), multiple front-ends, multiple phone sets, improved multipass decoding, and increasing the acoustic training data to 430 h, resulting in a WER of 21% (RT03 system, LIMSI component only).

A main factor for the RT04 evaluation was the availability of 2300 h of CTS data, mostly from the Fisher collection. Since the RT04 test data was also from the Fisher collection, a new set of development data (Dev04) was used to measure progress. Table VIII summarizes the main improvements in the LIMSI CTS system from RT03. An absolute error reduction of 1.7% was due to improved acoustic modeling by incorporating feature optimization (MLLT) and speaker adaptive training (SAT). An overall improvement of about 2.5% was obtained using the Fisher data after training better (and larger) acoustic (up to 1 million Gaussians) and language models, and updating the dictionary. Modifications to and incorporating acoustic model adaptation in a fast decode led to a gain of 0.4% while reducing the computation time by a factor of 6. Experiments using multiple phone sets gave an additional small improvement (see Section IV-E).

The advances made for the CTS task were ported to the BN task. The training techniques (MLLT, SAT, MMI), improved adaptation (CMLLR, MLLR), neural network language modeling, and consensus decoding with pronunciation probabilities were found to carry over. The only exception was VTLN, which did not consistently improve BN performance. In addition lightly supervised training was used to increase the quantity of BN audio training data. Although, at LIMSI, we did not develop stand alone BN systems for the evaluations, we estimate the annual performance improvements to be on the order of 20% relative while respecting the decoding time limitations (under 10 $\times$ RT).

TABLE VIII
SUMMARY OF IMPROVEMENTS TO THE LIMSI CTS COMPONENT SYSTEM FROM RT03. ABSOLUTE WER REDUCTIONS ON THE DEV04 SET AND OVERALL RELATIVE WORD ERROR REDUCTION

Improvement details	%WER red
Speaker adaptive training	0.9%
MLLT	0.8%
Improved models with Fisher data (LM, large AM, lexicon)	2.5%
Better and faster decoding with AM adaptation with factor of 6 speed-up	0.4%
Multiple phone sets modeling	0.4-0.7%
Overall relative error reduction without sys. combination	23%

B. Audio Segmentation and Acoustic Modeling

The LIMSI segmentation and clustering for BN is based on an audio stream mixture model [23], [25]. First, the nonspeech segments are detected and rejected using GMMs representing speech, speech over music, noisy speech, pure-music and other background conditions. An iterative maximum-likelihood segmentation/clustering procedure is then applied to the speech segments. The result of the procedure is a sequence of nonoverlapping segments with their associated segment cluster labels. The objective function is the GMM log-likelihood penalized by the number of segments and the number of clusters, appropriately weighted. Four sets of GMMs are then used to identify telephone segments and the speaker gender.

The acoustic models for BN and CTS use the same model topology and are constructed in a similar manner (the VTLN step is skipped for BN), depicted in Fig. 1. Over the duration of the EARS program, different acoustic feature vectors and phone sets have been explored, with two aims in mind: optimizing model accuracy for a given model set and developing acoustic models that combine well for within and cross-site adaptation and combination.

Each context-dependent phone model is a tied-state, left-to-right CD-HMM with Gaussian mixture observation densities. The acoustic feature vector has 39 components comprised of 12 cepstrum coefficients and the log energy (estimated on a 0–8 kHz band (or 0–3.5 kHz for telephone data), along with the first- and second-order derivatives. Two sets of gender-dependent, position-dependent triphones are estimated using MAP adaptation of SI seed models (for each bandwidth for BN). The triphone-based context-dependent phone models are word-independent but word position dependent. The first decoding pass uses a small set of acoustic models with about 5000 contexts and tied states. Larger sets of acoustic models covering 30 k–40 k phone contexts represented with a total of 11.5 k–30 k states are used in the latter decoding passes. State-tying is carried out via divisive decision tree clustering, constructing one tree for each state position of each phone so as to maximize the likelihood of the training data using single Gaussian state models, penalized by the number of tied-states [25]. There are about 150

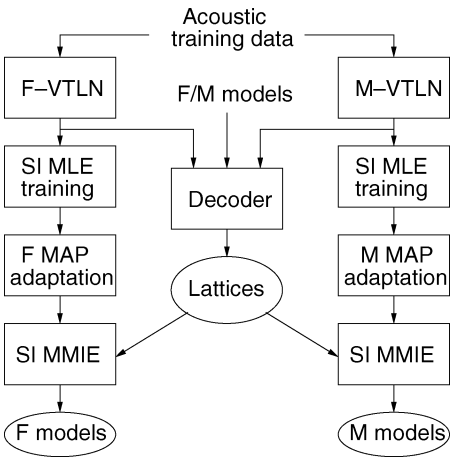


Fig. 1. LMSI CTS acoustic model training procedure.

questions concerning the phone position, the distinctive features (and identities) of the phone and the neighboring phones.

VTLN is a simple speaker normalization at the front-end level which performs a frequency warping to account for differences in vocal tract length, where the appropriate warping factor is chosen from a set of candidate values by maximizing the test data likelihood based on a first decoding pass transcription and some acoustic models (some sites, e.g., BBN, use GMMs with no need for transcriptions). Following [26], the Mel power spectrum is computed with a VTLN warped filter bank using a piecewise linear scaling function. We found the classical maximum-likelihood estimation procedure to be unsatisfying, as iterative estimates on the training data did not converge properly, even though a significant word error reduction on conversational speech was obtained. This problem can be attributed to the fact that the VTLN Jacobian is simply ignored during the ML estimation, although the normalization of the feature variances should largely compensate for this effect. Properly compensating the VTLN Jacobian would require building models for each possible warping value and would double the computation time to estimate the warping factors, we therefore investigated changing the procedure to avoid the Jacobian compensation.

The VTLN warping factors are still estimated for each conversation side by aligning the audio segments with their word level transcription for a range of warping factors (between 0.8 and 1.25), but we use single-Gaussian gender-dependent models to determine the ML warping factor. By using gender-dependent models (as proposed in [27]) the warping factor histogram becomes unimodal and is significantly more compact. This effect and the use of single Gaussian models (as proposed in [28]) significantly reduces the need for Jacobian compensation and makes the estimation procedure very stable, reducing the absolute WER by 0.4% after adaptation [29].

C. Training on Large BN Data Sets

One of the limitations in obtaining acoustic model training data is the high cost of producing manual transcriptions. Since several hundred hours of untranscribed audio data were available from the TDT corpora for which closed captions were also available, we used a semiautomatic approach [3] to generate training transcripts for the TDT4 portion of the data. The basic idea is to use a speech recognizer to generate a hypothesized transcription which is aligned with the closed captions. When the recognizer output and the closed captions agree, it is assumed that these words do not need to be corrected. If there is a disagreement, the recognizer output is given first (if there is one) followed by the closed caption text given in parentheses (all in uppercase letters).

Caption Correct	*Recognizer Correct*
IN (AND) ⇒ and	DOT (POINT) ⇒ dot
AND (IN) ⇒ in	DOCTOR (DR) ⇒ doctor
ALIVE (A LIVE) ⇒ a live	A (ONE) ⇒ a
BUILD (BUILT) ⇒ built	THE SECOND (TWO) ⇒ the second
FIFTEEN (FIFTY) ⇒ fifty	million (DOLLARS) ⇒ million
FOR (FOUR) ⇒ four	billion (DOLLARS) ⇒ billion
TO (TWO) ⇒ two	thousand (DOLLARS) ⇒ thousand

Fig. 2. Sample rewrite rules to correct the alignment of the closed captions with the recognizer hypotheses. On the left, the rules correct recognition errors (the caption is correct), and on the right, the recognizer is correct.

In our first usage of the TDT4 data for acoustic training, only audio segments where the hypothesized automatic transcription "agreed" with the associated aligned closed captions were used. Agreement means that the word error between the hypothesis and the caption was under 20% after applying some normalization rules to account for frequent, but inconsequential, errors. About 120 rules were used to correct some errors in the merged transcriptions after alignment. Fig. 2 shows two sets of example rules, one choosing the closed caption (within parentheses) as the correct answer, and the second set choosing the automatic transcription. The left part of the figure shows rules to correct some common recognition errors, whereas the rules on the right correspond to differences in the spoken and written forms. These rules are used to automatically decide (for each word sequence in uppercase) between the recognizer output and the closed caption, in order to get a single transcription usable for acoustic model training.

An alternative method for lightly supervised acoustic model training was also explored using consensus networks to provide more flexibility in aligning the system word lattice with the associated closed captions, thus potentially keeping additional training data [30].

D. Neural Network Language Model

Connectionist LMs [31], [32] have been explored for both the BN and CTS tasks. The basic idea is to project the word indices onto a continuous space and to use a probability estimator operating on this space. Both tasks are performed by a neural network. This is still an n-gram approach, but the n-gram LM probabilities are "interpolated" for any possible context of length $n-1$ instead of backing-off to shorter contexts. Since the resulting probability densities are continuous functions of the word representation, better generalization to unknown n-grams can be expected.

For BN, the neural network LM was trained on a subset of about 27 M words of data (BN transcriptions, TDT2, TDT3, and TDT4 closed captions and four months of CNN transcripts from

2001). The neural network LM is interpolated with a 4-gram backoff LM built by merging nine component models trained on, in addition to the above materials, commercially produced BN transcripts (260 M words); archived CNN web transcripts (112 M) and newspaper texts (1463 M words). Lattice rescoring with this NN LM is done in about 0.1 $\times$ RT. The perplexity of the Dev04 data is 109.9 for the 4-gram back-off LM alone and 105.4 when interpolated with the neural net LM. This results in an absolute word error reduction of 0.3% for the LIMSI components used in integrated BBN/LIMSI systems.

LIMSI has been using the neural network LM for CTS since the NIST RT02 evaluation. Although the word error rate of the complete system has decreased from 24% to under 19% due to the other improvements in the models and to the decoding procedure, the neural network LM always achieved a consistent word error reduction of about 0.5% absolute with respect to a carefully tuned 4-gram back-off LM [32]. In the final integrated RT04 BBN/LIMSI system for CTS described next, all three LIMSI components use the neural network LM. The neural network LM was trained on all of the CTS acoustic training data transcripts (27 M words) and interpolated with a 4-gram back-off LM trained on all the available data.

E. Alternative Phone Units

In an attempt to model some of the different speaking styles found in conversational telephone speech, we explored the use of pronunciation variants with probabilities, alternate phone sets, and syllable-position-dependent phone models. In addition to the standard 48 phone set used in the LIMSI CTS and BN systems [33], two of the alternate sets change the number of phones in a word. The reduced phone set [34], splits some of the complex phones into a sequence of two phones [35], thereby increasing the number of phones per word and potentially better matching slow speech. In the extended phone set, some selected phone sequences are mapped into a single unit in an attempt to better model heavily coarticulated and fast speech. These pseudophones can represent consonant clusters, vowel–liquid or vowel–nasal sequences. The expanded phone set leaves the number of phones unchanged, introducing syllable-position-dependent models for some phones which may have significantly different realizations in different syllable positions. Each model set has about 30 k tied states, with 32 Gaussians per state. The models based on the extended phone set are somewhat larger, with around 50 k states.

The recognition results on the CTS Dev04 data without MMI training are given in Table IX. The first three decodes use only the original phone set and result in a 17.5% word error rate after two passes of MLLR adaptation using two regression classes (speech and nonspeech) and four phonemic regression classes, respectively. The word error rates with the other model sets are given in the lower part of the table. These decodes are preceded by a four class MLLR adaptation with the same second-pass hypothesis. Comparable results are obtained for the three model sets, even though the extended set appears to have the highest error rate (17.8%).

No phone set was found to perform best for a majority of speakers. The standard and reduced phone sets each were best for 1/3 of the speakers, with the remainder divided between the expanded and extended phone sets. Listening to portions of the data from the speakers who had the lowest word error rates with the extended phone set, it appears that most of these speakers have a casual speaking style, with a tendency to slur some of their words. Despite our expectations that the reduced phone set would favor slow speakers, no correlation was found with a global estimate of the speaking rate in words per minute. Combining the models outputs with Recognizer Output Voting Error Reduction (ROVER) [36] reduces the WER to 16.8% showing the models to be somewhat complementary. The gain of the combined result was quite large for some speakers (4% absolute) with no large loss for any speaker [35], and with no notable improvement for speakers with slow or fast speech. Models based on the original and reduced phone sets were used in the LIMSI components in the combined BBN/LIMSI systems.

TABLE IX
WORD ERROR RATES ON THE CTS DEV04 DATA FOR THE FOUR PHONE SETS

Decode	*WER*
Unadapted fast decode	23.4
2 class adaptation	17.8
4 class adaptation	17.5
Reduced set, 4cl adapt.	17.3
Expanded set, 4cl adapt.	17.6
Extended set, 4cl adapt.	17.8
Combination	16.8

F. Decoding Architecture

Decoding is usually carried out in multiple passes for both the CTS and BN tasks where the hypothesis of one pass is used by the next pass for acoustic model adaptation. For each decoding pass, the acoustic models are first adapted using both the CMLLR and MLLR adaptation methods. MLLR adaptation relies on a tree organization of the tied states to create the regression classes as a function of the available data. This tree is built using a full covariance model set with one Gaussian per state. Then, a word lattice is produced for each speech segment using a dynamic network decoder with a 2-gram or a 3-gram language model. This word lattice is rescored with a 4-gram neural network language model and converted into a confusion network (using the pronunciation probabilities) by iteratively merging lattice vertices and splitting lattice edges until a linear graph is obtained. This procedure gives comparable results to the edge clustering algorithm proposed in [37], but appears to be significantly faster for large lattices. The words with the highest posterior in each confusion set are hypothesized along with their posterior probabilities.

For the CTS data, the first hypothesis is also used to estimate the VTLN warp factors for each conversation side. When the computation time budget is limited (cf. Section VII), Gaussian short lists and tight pruning thresholds are used to keep decoding time under 3 $\times$ RT.

V. NOTEWORTHY RESEARCH IDEAS

Besides the techniques described in the previous sections, BBN and LIMSI explored several other ideas that seemed promising, but did not become part of the final evaluation

systems because they did not improve recognition accuracy within the allotted time frame.

One such idea was the discriminative initialization of Gaussians in the HMM. In current state of the art systems, discriminative HMM training takes place after the Gaussians have been already estimated via several iterations of ML training. Under ML estimation, Gaussians in a given HMM state are positioned such that they cover the acoustic data aligned to that state. The allocation of Gaussians is performed independently for each HMM state, so no particular attention is devoted to optimizing the decision boundary between different HMM states. Moreover, although significant improvements in the ML criterion can be obtained by increasing the number of mixture components for states with lots of data, this gain typically comes from more detailed modeling of the interior region of the state underlying distributions, which may not be that useful from a discriminative point of view. In fact, having these extra components in the state mixtures could lead to overfitting after discriminative training, causing poor generalization on unseen data. Normandin [38] showed that by performing the Gaussian splitting initialization of state mixture distributions based on MMI, rather than ML, significant improvements in recognition accuracy can be achieved. A similar initialization procedure was used in the BBN system, but although the MMI criterion improved significantly on the training data, the WER on the unseen test set degraded. At that time, we suspected that the degradation could be explained by MMI's weak correlation with WER. Indeed, measurements on the training data showed that improvements in the MMI criterion did not always correspond to improvements in the WER. We believe that better results can be obtained with MPE, due to its strong correlation with WER.

Another area of exploration was the more accurate modeling of speech trajectories. The "conditional independence assumption" is often presented in the literature as a fundamental weakness of the HMM. More specifically, the assumption is that observation probabilities at a given HMM state depend only on that state and not on previous states or observations. This is clearly untrue for speech observations, so researchers have tried for many years to overcome this limitation of the HMM through a variety of methods. The easiest and most successful technique thus far has been to incorporate acoustic context information, both through the use of context-dependent HMM states, as well as by extending the observation vector to include differentials of base energy and cepstral coefficients computed over a window of time (typically 70–90 ms). Other techniques have been the so called "segmental models," which attempt to model explicitly the dependence of observations within a phoneme or subphoneme segment. As a compromise between the standard HMM and a segmental model, we explored the "convolutional trajectory model," which assumes that the observations within a phoneme segment are formed by adding the outputs of two independent stochastic processes: a segment-based process, that generates smooth phoneme trajectories, and a frame-based process, that generates small perturbations. In the convolutional model, the first process is modeled by a polynomial trajectory model, as in [39], while the residual is modeled by an HMM. The overall segment likelihood is the convolution of the (hidden) trajectory likelihood and the HMM likelihood of the residual between the trajectory and the actual observations. In the relatively short amount of time devoted to this research idea, we were able to develop a convolutional model that matched the standard HMM in terms of recognition accuracy performance on English CTS data. We still believe that there is promise in this research direction.

Aside from the aforementioned ideas, there were a few that were considered, but not tried, mostly due to limited amount of time. One such idea was the use of discriminative criteria in the decision tree-based HMM state clustering. Most systems today use a decision tree to group the HMM states into clusters for more robust parameter estimation. States at a given node (cluster) in the tree are divided into subclusters by asking binary linguistic questions about the phonetic context of the states; the question that splits the node with the largest increase in likelihood is selected, and the same process is repeated recursively on the descendant nodes, until a stopping criterion is met. We have found that the growing of the decision tree can be modified to allow both node splits and merges, resulting in significant likelihood increase, but with no improvement in recognition accuracy. This can be attributed to the fact that maximizing likelihood does not always guarantee improved performance on unseen data. Using a discriminative criterion, such as MPE, might provide better results.

The discriminative training of language models was another area of investigation that was in our plans but never realized. This is a particularly challenging area, because in standard n-gram-based LMs, the model parameters are not shared and there are tens or hundreds of millions of n-grams that need to be estimated. If a small amount of training data is used for discriminative training, such as the acoustic training data, overfitting is bound to occur. For this reason, the idea of discriminative training is more appealing in the case of the neural network LM, where the number of independent parameters is much smaller compared to a standard LM, and the mapping of the n-grams to a continuous space enables the exploration of gradient-based techniques.

VI. Integration of BBN and LIMSI Components

ROVER has been demonstrated as an effective method for combining different systems to achieve a WER that is significantly better than the result obtained by any of the individual systems. ROVER requires running two or more systems independently and then combining their outputs using voting. Due to the compute constraints enforced under the EARS program, combining a large number of different systems using ROVER was quite challenging. Therefore, we explored novel combination strategies to effectively combine multiple BBN and LIMSI systems.

A. Cascade versus ROVER

We explored most of the combination strategies by combining English CTS systems from BBN and LIMSI. As shown in Fig. 3(a), the baseline systems from both sites consisted of three recognition stages. The final WERs on the Eval01 test set of BBN's and LIMSI's (stand-alone) systems are 21.6% (running at 15 $\times$ RT) and 21.1% (at 20 $\times$ RT), respectively.

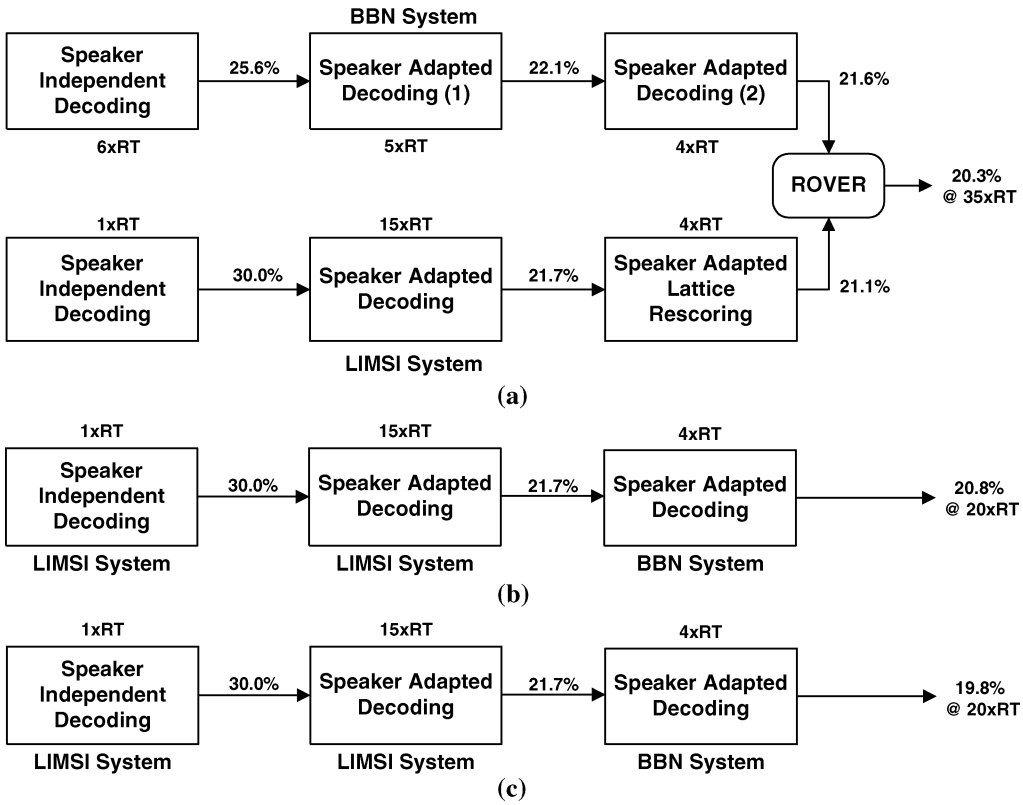


Fig. 3. Comparing Cascade and ROVER combination. (a) ROVER. (b) Cascade using two regression classes in the adaptation of the BBN system. (c) Cascade using eight regression classes in BBN adaptation.

Compute time was measured on a 2.8-GHz Pentium 4 Xeon CPU. The ROVER of these two results produced a WER of 20.3% that requires a total compute time of 35 $\times$ RT.

In order to reduce the computation, a "cascade" configuration was explored. As shown in Fig. 3(b), (c), instead of the "parallel processing" nature of ROVER, the decoding process is *sequential* in the "cascade" architecture. The BBN system took care of the very last decoding stage after adapting the acoustic models to the hypotheses generated by the second stage of the LIMSI system. When using only two regression classes during the adaptation of the acoustic models, we obtained a WER of 20.8%, while the overall running time was only 20 $\times$ RT as shown in Fig. 3(b).

We have found that cross-site adaptation, i.e., adaptation of one site's models with supervision based on the hypotheses generated by another site's system, typically requires larger number of regression classes. In fact, when using eight regression classes, the WER was reduced to 19.8% as shown in Fig. 3(c). It was clear that the "cascade" configuration outperformed the ROVER combination in both WERs (19.8% versus 20.3%) and compute time (20 $\times$ RT versus 35 $\times$ RT).

B. Combining Cascade and ROVER

In Section VI-A, the effectiveness of cascading recognition stages from different systems to exploit the gain from combining different systems without a significant increase in the compute time was demonstrated. Since we are "cross-adapting" one system to another in the cascade configuration, we wanted to evaluate whether performing a ROVER on the cross-adapted recognition hypothesis with the hypothesis used for adaptation results in any additional gain. This is depicted in Fig. 4. It is interesting to see that ROVER produced further improvement only in the case of using two regression classes for cross-adaptation (20.2% versus 20.8%).

C. Lattice versus Full Decode

Acoustic rescoring of word lattices is often used internally in a multiple-pass single site system to speed-up the decoding without a significant increase in the word error rate. The same approach can be envisioned for system combination.

A problem posed by cross-system lattice rescoring is that the lattices from one system must be transformed to be compatible with the vocabulary of the other system. One of the main factors that affects the recognizer vocabularies is the word tokenization used by each system, in particular with regards to the use of multiword sequences (often referred to as compound words) in order to improve better model reduced pronunciations and common contracted forms. To handle such differences, system specific sets of rules need to be developed to convert the lattices properly. Another issue with lattice rescoring is that the second decode is restricted to what the first system already explored. While this is an advantage for speedup, it may also reduce the possible gain of combination by making the second decode quite dependent on the first one. With the cascade style combination, there is even more dependency between the two decodes since the acoustic models used in the second decode are adapted by making use of the first pass hypotheses.

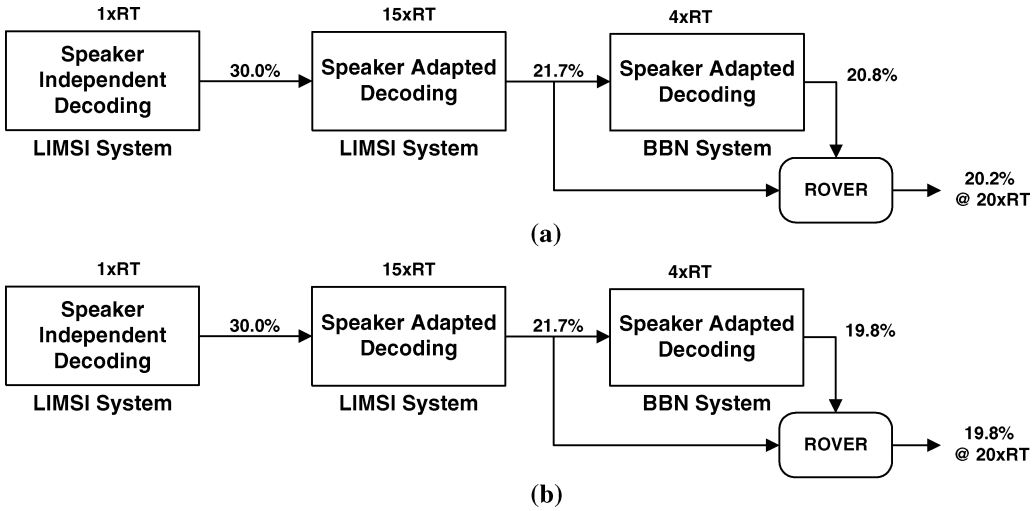


Fig. 4. Combination of Cascade and ROVER. (a) Two regression classes used for adapting the BBN system. (b) Eight regression classes used in BBN adaptation.

An alternative solution for cross-site system combination is to carry out a full decode after adaptation. This approach has the advantage of keeping the two decodes more independent, but it is only viable if very fast decoding techniques are used, which usually results in higher word error rates.

We carried out a series of experiments to compare these two solutions and found that cross-site adaptation with a full decode was both simpler and more efficient solution than lattice rescoring. This solution was therefore adopted for the RT04 evaluation.

D. Generic System Architecture

In Fig. 5 we present two different generic architectures for combining system components. In the first architecture, recognition hypotheses from multiple independent systems are combined using ROVER, then the ROVER output is used to adapt the same or a different set of systems. The adapted systems are used for another recognition pass. The compute requirements of this architecture are significantly higher than that of a cascade configuration because it requires running multiple independent systems.

We used the combination architecture described in Fig. 5(a) for the RT03 BBN/LIMSI English CTS evaluation system [40]. Three systems from BBN and two systems from LIMSI were first run independently. The recognition outputs of these five systems were combined using ROVER. The ROVER hypothesis was then used to adapt the acoustic models of the same five systems, and the adapted models were used in another recognition pass. The recognition hypotheses from the five adapted decodings were combined to generate the final output. Since the RT03 English CTS condition did not impose any constraints on the compute time, we were able to incorporate such large number of recognition stages. Unfortunately, that was not the case for the RT04 Evaluation. The RT04 English CTS condition limited the overall system's runtime at 20 $\times$ RT!

The second architecture is a combination of Cascade and ROVER based on the results presented above, where multiple systems are run sequentially. As shown in Fig. 5, the output of

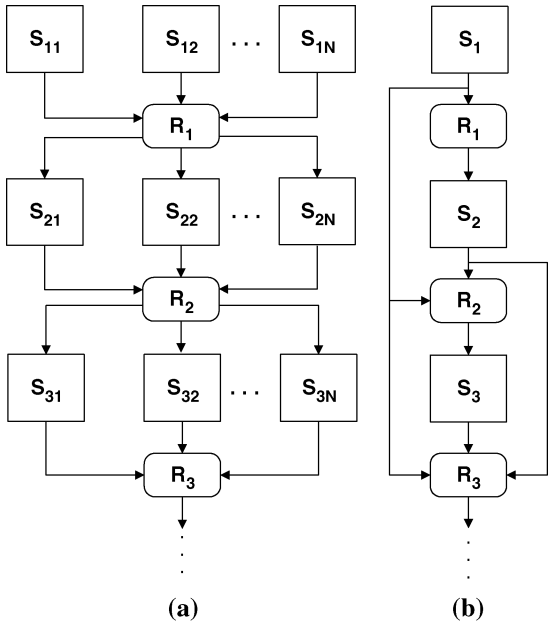


Fig. 5. Proposed architecture for combining multiple systems. The ROVER/Cascade system in (b) is designed for fast combination, whereas (a) does take compute time into consideration.

the system S_i or the ROVER hypothesis R_i can be used to adapt system S_{i+1}, where

$$R_i = \text{ROVER}(S_1, \ldots, S_i). \tag{1}$$

The Cascade/ROVER combination architecture was first used in the RT03 BBN/LIMSI English BN system [40]. In that system, we cascaded two recognition stages from LIMSI and one recognition stage from BBN. Finally, ROVER was used to combine outputs from BBN and LIMSI recognitions to achieve a WER that was better than ROVER of independent systems from each site and still satisfied the 10 $\times$ RT compute constraints.

Since compute requirements were even more challenging for the RT04 evaluations, the Cascade/ROVER combination described in Fig. 5(b) was used in the RT04 BBN/LIMSI systems for both domains.

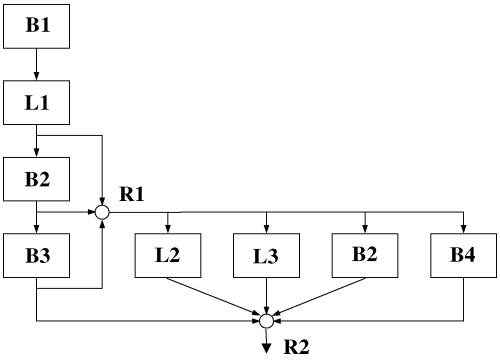


Fig. 6. 2004 BBN/LIMSI CTS system architecture.

VII. Development and Evaluation Results

A. The 2004 BBN/LIMSI CTS System Results

The block diagram of the 2004 BBN/LIMSI CTS tightly integrated system is depicted in Fig. 6. Systems from BBN are denoted with prefix "**B**" and those from LIMSI with prefix "**L**." An incoming arrow into a system indicates that the system's models are adapted to the previous result before decoding. Multiple incoming arrows into a small circle indicate that the results are combined using ROVER to produce a new hypothesis. It is worth mentioning that this configuration is a "cross-over" between the two forms of system combination depicted in Fig. 5. Key characteristics of each system are tabulated as follows.

B1 BBN PLP long span held-out MPE models.
B2 BBN PLP derivative MPE models.
B3 BBN PLP long span MPE models.
B4 BBN MFCC long span MPE models.
L1 LIMSI PLP GD MMI models.
L2 similar to **L1** with a reduced phone set.
L3 LIMSI PLP GI-MAP MMI models.

First, the waveforms were segmented using the BBN CTS segmenter described earlier. System B1 ran in slightly over real-time. Then B1's hypothesis was used to adapt L1's models with four fixed regression classes. System L1 decoded, using the same segmentation, in about 5 × RT to generate lattices and a new hypothesis. Next, B2's models were adapted to L1's hypothesis using a maximum of eight regression classes. System B2 then ran in about 3 × RT to generate a hypothesis to adapt B3's models, again using a maximum of eight regression classes. System B3 then ran a *full* three-pass decoding at 2.5 × RT and also saved the fast-match information (aka the reduced search space) for later *partial* two-pass decodings. Hypotheses from systems L1, B2, and B3 were combined using ROVER to produce hypothesis R1, which was used to adapt the models of systems L2, L3, B2, and B4 with a maximum of 16 regression classes. Systems L2 and L3 rescored L1's lattices while systems B2 and B4 performed a *partial* decoding on B3's reduced search space. The lattice rescoring took about 1.2 × RT and the partial decoding took 2.1 × RT. Finally, hypotheses from systems B3, L2, L3, B2, and B4 were combined to produce the final hypothesis R2.

Table X summarizes the WERs and real-time factors (× RT) for each decoding stage on both the CTS Dev04 and Eval04 test sets. The notation in the table shows the path producing the output, thus the name of the system includes the name of the preceding system, plus the new system that was run (for example, B1-L1-B2 indicates a system that first ran B1, then adapted, then L1, then adapted, then B2). Overall, for the CTS Dev04 and Eval04 test sets, the combined BBN/LIMSI CTS system performed at 13.4% WER in 18.5 × RT and at 16.0% WER in 18.0 × RT, respectively.

TABLE X
WER and Run-Time on the CTS Dev04 and Eval04 Sets From Each Stage of the 2004 BBN/LIMSI CTS 20 × RT System

System	Dev04		Eval04	
	%WER	xRT	%WER	xRT
B1	18.0	1.3	21.0	1.2
B1-L1	15.5	4.8	18.3	4.6
B1-L1-B2	14.4	3.1	16.9	3.0
B1-L1-B2-B3	14.2	2.6	16.7	2.5
R1	13.8	0.0	16.2	0.0
R1-L2	14.5	1.2	16.9	1.2
R1-L3	14.6	1.2	17.1	1.3
R1-B2	14.2	2.1	16.4	2.1
R1-B4	14.0	2.1	16.3	2.0
R2	13.4	0.0	16.0	0.0
Overall	13.4	18.5	16.0	18.0

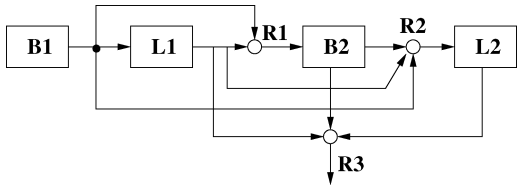


Fig. 7. 2004 BBNLIMSI BN system architecture.

B. 2004 BBN/LIMSI BN System Results

Similar to the CTS system, the 2004 BBN/LIMSI BN system also used both cross-site adaptation and ROVER for system combination. The system structure, slightly different from the CTS system, is depicted in Fig. 7. Key characteristics of each system are tabulated as follows.

B1 BBN PLP MMI SI and SAT system.
B2 BBN PLP MMI SAT system.
L1 LIMSI PLP GI-MAP MMI system.
L2 similar to L1 but with a reduced phone set.

The WERs and real-time factors on both the BN Dev04 and Eval04 test sets are listed in Table XI. Specifically, on the development test set Dev04, in the very first step, system B1 generated a hypothesis with an 11.0% error rate at less than 3 × RT using both unadapted and adapted decoding. Then, system L1, after adapting to B1's hypothesis, redecoded and produced a new hypothesis with 10.1% error rate. In contrast to the combined CTS system, BBN and LIMSI did not share the same audio segmentation. A ROVER of the hypotheses of systems B1 and L1 provided a hypothesis of 9.8% WER. System B2

The BBN compute platform is an Intel Xeon (3.4 GHz, 8 GB RAM) running Linux RedHat 7.3, with hyperthreading. At LIMSI the compute platform is an Intel Pentium 4 Extreme Edition (3.2 GHz, 4 GB RAM) running Fedora Core 2 with hyperthreading. To take advantage of hyperthreading, the test data was divided into two sets that were processed simultaneously by two decoding processes.

TABLE XI
WER AND RUN-TIME ON THE BN DEV04 AND EVAL04 SETS FROM EACH STAGE OF THE 2004 BBNLIMSI BN 10 × RT SYSTEM

System	Dev04		Eval04	
	%WER	RTF	%WER	RTF
B1	11.0	2.6	14.4	2.7
B1-L1	10.1	2.7	13.6	3.0
R1	9.8	0.0	13.2	0.0
R1-B2	9.9	2.1	13.4	2.2
R2	9.5	0.0	12.8	0.0
R2-L2	9.9	1.8	13.5	1.9
R3	9.3	0.0	12.7	0.0
Overall	9.3	9.2	12.7	9.8

TABLE XII
COMPARISON BETWEEN THE RT02 AND RT04 SYSTEMS ON THE CTS AND BN PROGRESS TEST SETS

System	CTS		BN	
	%WER	RTF	%WER	RTF
RT02	27.8	263	18.0	10.0
RT03	17.5	486	12.3	17.3
RT04	13.5	18.3	9.5	9.3

then adapted to the ROVER's result using a maximum of 16 regression classes and redecoded to produce a 9.9% result. Combining the hypotheses of systems B1, L1, and B2 produced a new result of 9.5% error rate. This latest ROVER's result provided supervision for the second LIMSI system, L2, which, in turn, produced a result of also 9.9% error rate. The final ROVER of the hypotheses of systems L1, B2, and L2 produced the final result of 9.3% error rate at 9.2 × RT. On the BN Eval04 test set, a word error rate of 12.7% was obtained at 9.8 × RT.

C. Results on the EARS Progress Test Set

As mentioned earlier, two progress test sets, one for BN and the other for CTS, were designated by NIST to benchmark the system performance yearly. The results on these two sets are given in Table XII. The BBN RT02 CTS and BN systems achieved 27.8% WER on the CTS progress test set and 18.0% WER on the BN progress set. These two systems were chosen as baselines against which progress was measured in the EARS program. The WER and real-time factor (RTF) obtained with the RT03 and RT04 BBN/LIMSI systems are reported in Table XII along with the baseline BBN RT02 numbers. The improvement is very significant with WERs for the RT04 systems reduced to 13.5% for the CTS data and to 9.5% for the BN data. Compared to the RT02 baseline systems, the relative WER reduction is 51% on the CTS data and 47% on the BN data. Furthermore, the compute time of the CTS system was also dramatically reduced. These results exceeded the EARS targets both in terms of recognition performance and decoding time constraints.

VIII. CONCLUSION

This paper has described the combined BBN/LIMSI system and the component systems developed at each site as part of the DARPA EARS program. Large word error rate reductions are reported for both the CTS and BN tasks, obtained by improving acoustic and language models, developing new training methods, exploiting unannotated training data, and developing innovative approaches for system combination.

By providing significantly larger amounts of audio and textual training materials, along with regular performance benchmarks, the program has fostered research in many directions and substantially improved the state-of-the-art in transcription of broadcast news data and conversational telephone speech in English, Arabic, and Mandarin.

The availability of the data, as well as the tools developed to process them, will enable numerous corpus based studies in the future. We believe that adding more training data will continue to help in improving recognition accuracy. However, the effect of additional data is expected to be limited unless a substantial amount is provided (e.g., tens of thousands of hours of speech). Given that, our focus is still concentrated on algorithmic and modeling improvements. Potential candidates for future research are as follows:

- better feature extraction (long span features, discriminative feature projections);
- improved speaker adaptive training;
- more detailed covariance modeling;
- adaptive pronunciation and language modeling;
- automatic training of complementary models to aid in system combination.

Our experience with the non-English languages addressed in the EARS program is that the same basic technologies and development strategies appear to port well from one language to another. However, to obtain optimal performance, language specificities must be taken into account. It may be that as word error rates are lowered, the language-dependent issues will become more important, and language-specific knowledge will help to improve performance.

REFERENCES

[1] C. Cieri, D. Miller, and K. Walker, "From switchboard to fisher: telephone collection protocols, their uses and yields," in *Proc. Eurospeech*. Geneva, Switzerland, Sep. 2003, pp. 1597–1600.
[2] O. Kimball, C. Kao, R. Iyer, T. Arvizo, and J. Makhoul, "Using quick transcriptions to improve conversational speech models," in *Proc. Int. Conf. Spoken Language Process.*. Jeju Island, Korea, Sep. 2004, pp. 2265–2268.
[3] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 115–229, 2002.
[4] L. Nguyen and B. Xiang, "Light supervision in acoustic model training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, May 2004, pp. 185–188.
[5] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in *Proc. HLT/NAACL*, 2003, pp. 7–9.
[6] L. Nguyen and R. Schwartz, "Efficient 2-pass N-best decoder," in *Proc. Eurospeech*, Rhodes, Greece, Sep. 1997, pp. 167–170.
[7] ——, "Single-tree method for grammar-directed search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Mar. 1999, pp. 613–616.
[8] S. Matsoukas, T. Colthurst, O. Kimball, A. Solomonoff, and H. Gish, "The 2002 BBN byblos English LVCSR system," presented at the *DARPA Speech Recognition Workshop*, Vienna, VA, May 2002.
[9] A. Andreou, T. Kamm, and J. Cohen, "Experiments in Vocal tract normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.
[10] R. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, May 1998, pp. 661–664.

[11] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. Int. Conf. Spoken Language Processing*, Philadelphia, PA, Oct. 1996, pp. 1137–1140.
[12] L. Nguyen and S. Matsoukas *et al.*, "The 1999 BBN BYBLOS 10 × RT broadcast news transcription system," in *Proc. Speech Transcription Workshop*.
[13] J. Davenport, R. Schwartz, and L. Nguyen, "Toward a robust real-time decoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Mar. 1999, pp. 645–648.
[14] D. Liu and F. Kubala, "A cross-channel modeling approach for automatic segmentation of conversational telephone speech," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, St. Thomas, U.S Virgin Islands, Nov. 2003, pp. 333–336.
[15] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, pp. 75–98, 1998.
[16] S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, Nov. 2003, pp. 273–278.
[17] R. Prasad and S. Matsoukas *et al.*, "The 2004 BBN/LIMSI 20 × RT english conversational telephone speech system," in *Proc. Rich Transcription Workshop*, Palisades, NY, Nov. 2004.
[18] P. C. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *Proc. ISCA ITRW ASR*, 2000.
[19] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 105–108.
[20] D. Povey *et al.*, "EARS progress update," presented at the *EARS STT Meeting*, St. Thomas, U.S. Virgin Islands, Nov. 2003.
[21] B. Zhang and S. Matsoukas, "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, May 2005, pp. 925–928.
[22] D. B. Paul, "An investigation of Gaussian shortlists," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 1999, pp. 209–212.
[23] J. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Commun.*, vol. 37, no. 1–2, pp. 89–108, 2002.
[24] Y. Bengio and R. Ducharme, "A neural probabilistic language model," in *Advances in Neural Information Processing Systems (NIPS)*. San Mateo, CA: Morgan Kaufmann, 2001, vol. 13.
[25] J. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *Proc. Int. Conf. Spoken Language Process.*, vol. 4, Sydney, Australia, Dec. 1998, pp. 1335–1338.
[26] T. Hain, P. Woodland, T. Niesler, and E. Whittaker, "The 1998 HTK system for transcription of conversational telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, AZ, Mar. 1999, pp. 57–60.
[27] P. Dogin, A. El-Jaroudi, and J. Billa, "Parameter optimization for vocal tract length normalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, Jun. 2000, pp. 1767–1770.
[28] L. Welling, R. Haeb-Umbach, X. Aubert, and N. Haberland, "A study on speaker normalization using vocal tract normalization and speaker adaptive training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1998, pp. 797–800.
[29] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational telephone speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, China, Apr. 2003, pp. I-212–I-215.
[30] L. Chen, L. Lamel, and J.-L. Gauvain, "Lightly supervised acoustic model training using consensus networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, May 2004, pp. 189–192.
[31] H. Schwenk and J.-L. Gauvain, "Connectionist language modeling for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. I-765–I-768.
[32] ——, "Neural network language models for conversational speech recognition," in *Proc. Int. Conf. Spoken Language Process.*, Jeju Island, Korea, Oct. 2004, pp. 1215–1218.
[33] L. Lamel and G. Adda, "On designing pronunciation lexicons for large vocabulary, continuous speech recognition," in *Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, Oct. 1996, pp. 6–9.
[34] L. Lamel and J.-L. Gauvain, "Continuous speech recognition at LIMSI," in *Final review of the DARPA Artificial Neural Network Technology (ANNT) Speech Program*, Stanford, CA, Sep. 1992, pp. 59–64.
[35] ——, "Alternate phone models for CTS," in *Proc. Rich Transcription Workshop*, Palisades, NY, Nov. 2004.
[36] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Santa Barbara, QC, Canada, May 1997, pp. 347–354.
[37] L. Mangu, E. Brill, and A. Stolke, "Finding consensus among words: Lattice-based word error minimization," in *ISCA Eurospeech*, Budapest, Hungary, Sep. 1999, pp. 495–498.
[38] Y. Normandin, "Optimal splitting of HMM Gaussian mixture components with MMIE training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 1995, pp. 449–452.
[39] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 1993, pp. 447–450.
[40] R. Schwartz *et al.*, "Speech Recognition in Multiple Language and Domains: the 2003 BBN/LIMSI EARS System," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. III, Montreal, QC, Canada, May 2004, pp. 753–756.

Spyros Matsoukas (M'98) received the B.Sc. degree in computer engineering and informatics from the University of Patras, Patras, Greece, in 1994, and the M.Sc. degree in computer science from Northeastern University, Boston, MA, in 1996.

Since then, he has been with BBN Technologies, Cambridge, MA, where he is a Senior Scientist conducting research in speech recognition and machine translation.

Jean-Luc Gauvain (M'98) received the Ph.D. degree in electronics from the University of Paris XI, Paris, France, in 1982.

He is a Senior Researcher at the CNRS, Paris, where he is head of the LIMSI Spoken Language Processing Group. His primary research centers on large-vocabulary continuous speech recognition and audio indexing. His research interests also include conversational interfaces, speaker identification, language identification, and speech translation. He has participated in many speech-related projects both at the French National and European levels, as well as in the DARPA EARS and GALE programs. He has led the LIMSI participation in the DARPA/NIST organized evaluations since 1992, most recently for the transcription of broadcast news data and of conversational speech. He has over 220 publications.

Dr. Gauvain received the 1996 IEEE SPS Best Paper Award in Speech Processing and the 2004 ISCA Best Paper Award for a paper in the Speech Communication Journal. He was a member of the IEEE Signal Processing Society's Speech Technical Committee from 1998 to 2001, and is currently Coeditor-In-Chief of the *Speech Communication* journal.

Gilles Adda received the Engineering degree from the Ecole Centrale de Lyon, Lyon, France, in 1982, and the Docteur-Ingenieur degree in computer science from the Paris XI University, Paris, France, in 1987.

Since 1987, He has been a Research Engineer at LIMSI-CNRS, Paris, where he is responsible for the research activities on "linguistic models for spoken language." He has participated in a large number of European and National projects. His current interests are in linguistic models, corpus-based linguistics, evaluation and standards, and automatic speech recognition.

Dr. Adda is member of the ISCA Association and received, together with J. L. Gauvain and L. F. Lamel, the "ISCA Speech Communication Journal best paper award 2004."

Thomas Colthurst received the B.Sc. degree from Brown University, Providence, RI, in 1992, and the Ph.D. degree in mathematics from the Massachusetts Institute of Technology, Cambridge, in 1997.

Since then, he has worked on large-vocabulary conversational telephone speech recognition at BBN Technologies, Cambridge.

Chia-Lin Kao (M'03) received the S.M. degree in computer science from Harvard University, Cambridge, MA, in 1995.

She is a Scientist at BBN Technologies, Cambridge, working primarily on large-vocabulary speech recognition systems. She is currently involved in advanced speech encoding systems and has been with the company since 1995.

Owen Kimball (M'84) received the B.A. degree in mathematics from the University of Rochester, Rochester, NY, the M.S. degree in computer science from Northeastern University, Boston, MA, and the Ph.D. degree in electrical engineering from Boston University, Boston.

Since 1982, he has been with BBN Technologies, Cambridge, MA, where he is a Senior Scientist in the Speech and Language Processing Department. His research interests include speech recognition, speaker verification, and information extraction.

Dr. Kimball received the 2000 Best Paper Award from the IEEE Signal Processing Society.

Lori Lamel (M'88) received the Ph.D. degree in electrical engineering and computer sciience from the Massachusetts Institute of Technology, Cambridge, in 1988

Since 1991, She has been a Senior Researcher in the Spoken Language Processing Group, LIMSI, CNRS, Paris, France. Her research activities include speech recognition, studies in acoustic-phonetics, lexical and phonological modeling, and spoken dialog systems. She has been a prime contributor to the LIMSI participations in DARPA benchmark evaluations and responsible for the American English pronunciation lexicon. She has been involved in many European projects, most recently the IP's Chil and TCStar.

Dr. Lamel is a member of the Speech Communication Editorial Board, the Interspeech International Advisory Council and the Advisory Committee of the AFCP. She was a member of the IEEE Signal Processing Society's Speech Technical Committee from 1994 to 1998 and the EU-NSF Working Group for "Spoken-Word Digital Audio Collections," She has over 190 reviewed publications and is corecipient of the 2004 ISCA Best Paper Award for a paper in the *Speech Communication* Journal.

Fabrice Lefevre received the degree in electrical engineering from ENSEA-Cergy and the Ph.D. degree in computer science from the University Pierre et Marie Curie, Paris VI, Paris, France, in 2000.

He was appointed an Assistant Professor position at the University of Orsay, Paris XI, in 2001 where he worked in the Spoken Language Processing Group at LIMSI-CNRS. He joined the University of Avignon in 2005, where he works in the Human–Machine Dialog Team at LIA. His primary research activities include automatic speech recognition, speech understanding, and spoken dialog systems. He was involved in several European projects (CORETEX, AMITIES and LUNA) and also in the US DARPA-funded EARS project. He participated in several international (NIST) and French (AUPELF, Technolangue) spoken language recognition and understanding system evaluation campaigns.

Dr. Lefevre is a member of the International Speech Communication Association. He is also cofounder of the French Spoken Communication Association (AFCP) and a member of its administrative board since 2001.

Jeff Z. Ma (M'01) received the B.Sc. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 1989, the M.Sc. degree in pattern recognition from the Chinese Academy of Sciences, Beijing, China in 1992, and the Ph.D. degree in computer engineering from University of Waterloo, Waterloo, ON, Canada, in 2000.

He joined BBN Technologies, Cambridge, MA, in 2000. Since then, he has been conducting research on various aspects of speech and language processing, including speech recognition, topic identification, automatic machine translation, and information extraction and distillation, and he has also helped to develop commercial speech products.

John Makhoul (S'64–M'70–SM'78–F'80) received the B.E. degree from the American University of Beirut, Beirut, Lebanon, the M.Sc. degree from The Ohio State University, Columbus, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, all in electrical engineering.

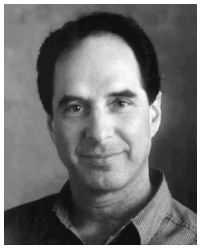
Since 1970 he has been with BBN Technologies, Cambridge, where he is a Chief Scientist working on various aspects of speech and language processing, including speech recognition, optical character recognition, language understanding, speech-to-speech translation, and human–machine interaction using voice. He is also an Adjunct Professor at Northeastern University, Boston, MA.

Dr. Makhoul has received several IEEE awards, including the IEEE Third Millennium Medal. He is a Fellow of the Acoustical Society of America.

Long Nguyen (M'93) is a Senior Scientist at BBN Technologies, Cambridge, MA, where he has been working on various areas of pattern recognition, including speech recognition, optical character recognition, and face identification since 1990.

Rohit Prasad (M'99) received the B.E. degree in electronics and communications engineering from Birla Institute of Technology, Mesra, India, in 1997, and the M.S. degree in electrical engineering from Illinois Institute of Technology (IIT), Chicago, in 1999.

He is a Scientist at BBN Technologies, Cambridge, MA. At IIT, he worked on predictive spectral quantization for low bit-rate speech coding. Since joining BBN in 1999, he has performed research and development in various areas including large-vocabulary speech recognition, speech-to-speech translation, optical character recognition, information retrieval, and high-speed topic classification.



Richard Schwartz has directed and been involved in several areas of speech research including continuous speech recognition, phonetic vocoding, narrowband speech transmission, speaker identification and verification, speech enhancement, and phonetic speech synthesis at BBN Technologies, Cambridge, MA. He has also been involved in research toward real-time, low-cost, high-performance speech recognition. He has also been involved in research toward statistical methods for understanding and has developed novel methods for topic classification, document retrieval, and a powerful statistical system for finding named entities. He has adapted speech recognition techniques for applications in handwriting recognition and language-independent optical character recognition. More recently, he has been working on automatic translation of text and speech.

Holger Schwenk (M'02) received the M.S. degree from the University of Karlsruhe, Karlsruhe, Germany, in 1992 and the PhD degree from the University Paris VI, Paris, France, in 1996, both in computer science.

He then did postdoctorate studies at the University of Montreal and at the International Computer Science Institute, Berkeley, CA. Since 1998, he has been an Assistant Professor at the University of Paris XI and he is a member of the Spoken Language Processing group at LIMSI. His research activities focus on new machine learning algorithms with application to human–machine communication, in particular, handwritten character recognition, large-vocabulary speech recognition, language modeling, and statistical machine translation. He has participated in several European- and DARPA-funded projects (CoreTex, Tc-Star, Chil, EARS, Gale). He has over 30 reviewed publications.

Dr. Schwenk is a member the International Speech Communication Association.

Bing Xiang (M'03) received the B.S. degree in electronics in 1995 and the M.E. degree in signal and information processing in 1998, both from Peking University, Beijing, China, and the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, in 2003.

Since 1994, he has worked on speech recognition and speaker recognition in various laboratories, including the IBM Thomas J. Watson Research Center, Yorktown Heights, NY. He was also an invited member of the 2002 Johns Hopkins CLSP summer workshop. In January 2003, he joined the Speech and Language Processing Department, BBN Technologies, Cambridge, MA. His research interests include large-vocabulary speech recognition, speaker recognition, machine translation, and statistical pattern recognition.

Dr. Xiang has been an active reviewer for IEEE journals and conferences.