30. Towards Superhuman Speech Recognition

M. Picheny, D. Nahamoo

After over 40 years of research, human speech recognition performance still substantially outstrips machine performance. Although enormous progress has been made, the ultimate goal of achieving or exceeding human performance – *superhuman* speech recognition – eludes us. On a more-prosaic level, many industrial concerns have been trying to make a go of various speech recognition businesses for many years, yet there is no clear *killer app* for speech. If the technology were as reliable as human perception, would such *killer apps* emerge?

Either way, there would be enormous value in producing a recognizer with superhuman capabilities. This chapter describes an ongoing research program at IBM that attempts to address achieving superhuman speech recognition performance in the context of the metric of word error rate. First, a multidomain conversational test set to drive the research program is described. Then, a series of human listening experiments and speech recognition experiments based on the test set is presented. Large improvements in recognition performance can be achieved through a combination of adaptation, discriminative training, a combination of knowledge sources, and simple addition of more data. Unfortunately, devising a set of informative listening tests synchronized with the multidomain test set proved to be more difficult than expected because of the highly informal nature of the underlying speech. The problems encountered in performing the listening tests are presented along with suggestions for future listening tests. The chapter concludes with a set of speculations on the best way for speech recognition research to proceed in the future in this area.

30.1 Current Status

After over 40 years of research, human speech recognition performance still substantially outstrips machine performance. Although enormous progress has been made, the ultimate goal of achieving or exceeding human performance – *superhuman* speech recognition – eludes us. In addition, it is fair to say that there have been no recent *breakthroughs* in the speech recognition area – progress over the last several years, though continual, has been evolutionary rather than revolutionary. Can we achieve levels of superhuman speech recognition in our lifetimes through evolutionary approaches, or do we need to make radical changes to our methodologies?

On a more-prosaic level, many industrial concerns have been trying to make a go of various speech recognition businesses for many years, yet there is no clear *killer app* for speech that would guarantee the huge revenue needed to justify the expense of investing in the technology. Is this because there really is no value in the technology, or that the technology just needs additional incremental improvements, or is this because the levels of accuracy must reach human performance or higher to be regarded as sufficiently reliable to serve as a user interface?

Either way, there would be enormous value in producing a recognizer with superhuman capabilities. This chapter describes an ongoing research program at IBM that attempts to address some of these questions, focusing on word error rate as a metric. It is recognized that this is only one of many dimensions in which speech recognition performance can be measured, and it is hoped that this chapter will trigger corresponding studies in related areas, such as concept and meaning extraction. First, we describe a multidomain conversational test set established to drive the research program (Sect. 30.2). Then, we describe a series of human listening experiments that attempt to determine the best set of research investments to achieve the goal of superhuman speech recognition (Sect. 30.3), a set of recognition experiments that begin to address methodologies for designing systems to achieve superhuman performance (Sect. 30.4), and a set of speculations on the best way for research to proceed in the future in this area (Sect. 30.5).

30.2 A Multidomain Conversational Test Set

We had a number of goals in mind when we designed the test set. First, the test set had to cover a reasonably broad range of conversational applications and contain data representing key challenges to reliable recognition including various forms of acoustic interference, speech from non-native speakers, and a large recognition vocabulary. Second, the test set had to include at least one component that is readily available to other researchers to facilitate comparisons between our recognizers and those developed externally. Third, the test set needed to be reasonably small, to facilitate rapid turnaround of experiments. For all experiments reported here, we used a test set composed of the following five parts.

swb98: The switchboard portion of the 1998 hub 5e evaluation set [30.1], consisting of 2 h of telephone-bandwidth (8 kHz) audio. The data were collected from two-person conversations between strangers on a preassigned topic. A variety of telephone channels and regional dialects are represented in the data.

mtg: An initial release of the Bmr007 meeting from the ICSI (International Computer Science Institute) meeting corpus [30.2], consisting of 95 minutes of audio. The data were collected from eight speakers wearing either lapel microphones or close-talking headsets. This meeting involved eight speakers: five native speakers of American English (two females and three males), and three non-native speakers (all males). Although the data is wide bandwidth (16 kHz), the primary challenge in this test set is the presence of background speech in many of the utterances. The crosstalk problem is especially severe for speakers recorded using lapel microphones.

cc1: 0.5 h of audio from a call center. The data are collected from customer-service representatives (CSRs) and customers calling with service requests or problems. The primary challenge in this test is acoustic interference: a combination of nonlinear distortion from speech compression, background noise from both the CSR and customer sides, and intermittent beeps on the channel, which are played to remind the customer that the call is being recorded. The data is telephony bandwidth but otherwise relatively quiet.

cc2: 0.5 h of audio from a second call center. The recordings are from a different center than the cc1 test set, but cover similar subject matter and have similar, poor acoustics. This data set has no information associating speakers with sets of utterances, which poses problems for speaker and channel adaptation. The data is telephony bandwidth but otherwise relatively quiet.

vm: Test data from the IBM voicemail corpus, consisting of 1 hour of audio. This material was previously reported on as the E-VM1 test set [30.3], and is a superset of the test data in the voicemail corpus part I and part II distributed by the LDC (Linguistic Data Consortium). Unlike the other tests, the voicemail data are conversational monologues. The acoustic quality of the data is generally quite high, although loud clicks caused by the speaker hanging up at the end of some messages can pose problems for feature normalization – especially

Table 30.1 Characteristics of the multidomain conversational test set used for listening and recognition experiments

Task	Number of hours	Number of segments
swb98	2.0	3500
mtg	1.5	2060
cc1	0.5	978
cc2	0.5	1033
vm	1.0	1033

normalization of c0 based on the maximum value of c0 within an utterance. This test set also has no information associating speakers with sets of utterances. The data is telephony bandwidth but otherwise relatively quiet.

The test sets were segmented into utterances suitable for recognition using a variety of means. The default LDC segmentation was used for the swb98 data. An automatic segmenter was used on the mtg data. The cc1 and cc2 data came in the form of calls and were automatically segmented into smaller units; the vm data came in the form of individual messages and was also automatically segmented into smaller units. A summary of the total number of hours and segments for each test set is given in Table 30.1.

30.3 Listening Experiments

Humans use whatever information is available to aid recognition performance. In a recent classic paper, *Lippmann* [30.4] compared machine and human recognition performance across a wide variety of stimuli. He demonstrated that human performance far exceeded machine performance with minimal linguistic information (digits, letters, nonsense sentences), minimal acoustic information (speech in noise), and various combinations of both (telephone conversations). Table 30.2 shows human versus machine performance for digits, letters, sentences from the *Wall Street Journal*, and telephony conversations. In all four cases, human performance was significantly better than machine performance, sometimes by more than an order of magnitude. Although these numbers were obtained some years ago, one would be hard pressed to argue that more than a factor of two improvement in recognition has occurred in the last 10 years, so the gap is still substantial.

Many studies have suggested that humans can accurately identify words with as little as two seconds of surrounding context [30.5, 6]. *Allen* [30.7] presents strong evidence that humans can highly accurately recognize phonemes without any linguistic context, and also suggests that the enormous robustness of human speech perception across degradations in channel conditions arises from the clever processing of many independent frequency bands in the auditory system. Internal experiments performed at IBM by Jelinek and his associates in the late 1980s suggest that human performance in word prediction from text (the *Shannon Game* [30.8]) is over three times better than the language models then (and unfortunately, still currently) used in speech recognition when humans are presented with full sentential context. No studies seem to exist that assess the relative importance of these information sources, so we initiated a series of tests to try to ascertain the performance of humans on the multidomain test set.

Table 30.2 Human versus machine recognition word error rates across a variety of tasks (after *Lippmann* [30.4])

Task	Machine performance %	Human performance %
Connected digits	0.72	0.009
Letters	5	1.6
Resource management	3.6	0.1
WSJ (Wall Street Journal)	7.2	0.9
Switchboard	43	4

30.3.1 Baseline Listening Tests

The first set of experiments attempted to obtain baseline performance figures on how well humans could actually recognize the multidomain test set. Random segments were chosen from the test set and played to two listeners over headphones. In one experiment, the listeners were allowed to listen only one time to the utterances, and in the second experiment, the listeners were allowed to hear the utterances multiple times. The utterances were randomized across the test set both in terms of domain and in terms of order. In addition, long utterances were segmented into no longer than 2–3 second chunks, as the memory load for utterances that were longer essentially made the single-pass task intractable.

The summary results are shown in Table 30.3. The error rates are much higher than those presented in [30.4] on similar data. In addition, no appreciable reduction in error rate is seen when the listener is allowed to listen multiple times to the test utterance. The error rates

Table 30.3 Summary of the multidomain test set base listening test results

Condition	Word error rate
Listen once	25.6
Listen multiple times	21.5

Table 30.4 Breakdown of base listening test results across corpora (listen-once case)

Corpus	Word error rate
swb98	25.1
mtg	28.1
cc1	19.6
cc2	32.2
vm	18.8

by corpus were broken down to see if there were any obvious dependencies on the domain materials. The breakdown is shown in Table 30.4. As can be seen, there is no clear dependency on any one particular corpus, though it is clear, for example, that the voicemail data is more comprehensible than the call center data, perhaps because the talkers know that they are leaving a message that is to be listened to after the fact.

Why is there such a discrepancy in the results here and the results reported in [30.4]? There are a number of possible causes. First, the earlier experiments were performed on longer segments of speech, listened to multiple times by each listener. Second, the earlier experiments may not have randomized the segments across speaker. Therefore, the listener may have been able to take advantage of both task adaptation and speaker/channel adaptation in the earlier results. In an attempt to try to tease these issues apart, we embarked upon a more-ambitious series of listening tests, described in the next subsection.

30.3.2 Listening Tests to Determine Knowledge Source Contributions

The processes that determine how humans recognize speech are still subjects of ongoing research and will remain so for some time to come. The goal of a speech recognition professional is not so much to understand how humans recognize speech but to try to utilize human strategies of speech recognition as a guide to improve the performance of today's speech recognition systems. As such, the next set of listening experiments attempt to determine what sources of knowledge humans can use in the context of performance of today's speech recognition systems.

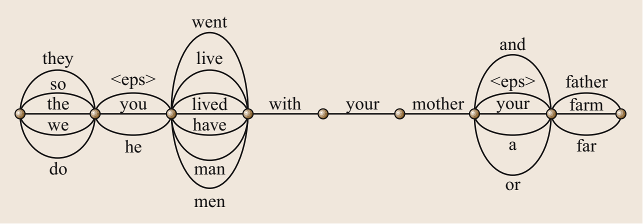


Fig. 30.1 A typical word confusion network, also known as a *sausage*

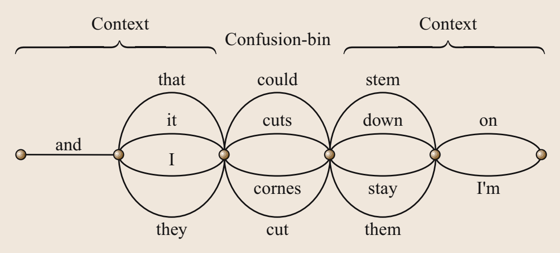


Fig. 30.2 Word recognition test for the *listening* condition depicting only the audio information and sausage structure presented to the subjects

Currently, one of the most interesting aspects of speech recognition systems is that even when they produce relatively high word error rates (e.g., 30%), it is possible for them to produce extremely compact representations of the search space in which the best possible, or oracle, error rate is very low (e.g., 10%). These representations take the form of a confusion network (typically referred to as a *sausage*), as illustrated in Fig. 30.1. In this example, the correct word sequence is *so you lived with your mother and father,* although *do you live with your mother and father* and many other sequences are also possibilities. (The word that was ranked highest by the recognizer appears at the top of each segment.)

In this set of listening experiments, sausage structures produced by a state-of-the-art recognizer were presented to human subjects, and they were asked to select the best word sequence either with accompanying audio (*listening*) or with accompanying long language model context (*comprehension*). The hope was to assess the relative contributions of acoustic and language information as possible information sources on top of a pre-existing speech recognition system.

More specifically, in the listening condition (Fig. 30.2) subjects were presented with audio containing two words to the left, the target word and two words to the right of the word target, and asked to identify the target center word. In the comprehension condition (Fig. 30.3), the subject was given five segments of text history and asked to determine the best path through the sausage network without corresponding audio. In these experiments, we used audio data from the RT03 evaluation data set from LDC [30.9] and from the

Table 30.5 Word and oracle recognition error rates for listening tests to determine relative knowledge source contributions

Corpus	Word error rate	Oracle error rate
RT03	29.2	8.0
MALACH	28.3	9.5

Table 30.6 Listener error rates when presented with audio context in the form of the surrounding words (listening) and word context in the form of sausages (comprehension)

Corpus	Base recognition word error rate	Listening	Comprehension
RT03	29.2	24.2	30.6
MALACH	28.3	27.3	32.0

MALACH corpus [30.10]. A summary of the word error rates and oracle error rates for the sausages are shown in Table 30.5. As can be seen, the oracle error rates are much lower than the word error rates, implying that, if there is useful information that humans can utilize in the surrounding context, error rates can be made to drop significantly.

Sausages as described above were presented to a set of 20 listeners in random order with respect to speaker, channel, and topic. Each listener saw or heard 100 sausages either under the listening or comprehension conditions. The results are shown in Table 30.6. As can be seen from the results, neither source of information dramatically improved the listener's ability to choose the correct path through the sausage data. The presence of acoustic information helped somewhat more than textual context, but not in some definitive sense. Subjectively speaking, the task was surprisingly difficult to perform – the conversations had significant sections of quite unintelligible murmuring which may not be critical for comprehension but when segments are selected completely at random may significantly affect the overall error rate. In retrospect, it might have been useful to include a condition in which both long-span acoustic and language model information was included to ensure that humans could perform the task with *infinite* context, and also have a contrastive condition in which the target speaker is fixed across a variety of utterances. There were issues associated with occasional misalignments of the target word text and the underlying audio and some segmentation artifacts, but these did not occur frequently enough to affect the error rates in a significant fashion.

"... spending and an indepted economy. But I have a really good friend who keeps trying to convince me that being in dept is a healthy economy. And I just do not see that. I l wish that uh ..."

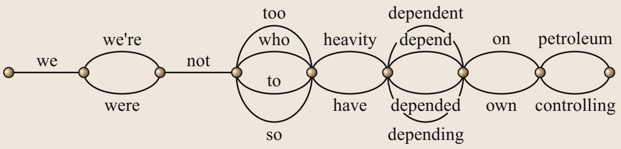


Fig. 30.3 Word recognition test for the comprehension condition depicting only the text information and the sausage structure presented to the subjects

What these initial experiments illustrate are some of the difficulties in performing human listening experiments as a guide for a technical research agenda in speech recognition. Unfortunately, we were unable to adequately tease out where research efforts should be focused to achieve the maximum benefit, and were left to rely on our technical intuition as to how best to proceed.

30.4 Recognition Experiments

To achieve superhuman speech recognition, it is useful to decompose speech recognition into a set of semi-independent tasks. This is shown in Table 30.7. The error rates are rough guides and should not be taken

Table 30.7 Recognition performance in terms of word error rates for different styles and types of speech

Task	Speech style	Target	Channel	Word error rate
Dictation	Well formed	Computer	Full BW (bandwidth)	< 4 [30.11]
Broadcast news	Usually well formed	Spontaneous	Audience	8.6 [30.12]
DARPA communicator	Spontaneous	Computer	Telephone BW	15.1 [30.13]
SWB	Spontaneous	Person	Telephone BW	15.2 [30.14]
Voicemail	Spontaneous	Person	Telephone BW	27.9 [30.3]
Meetings	Spontaneous	People	Far-field	50 [30.15]

literally. As can be seen, the main determinants of recognition difficulty are the mode of speech (is one talking to a machine or to a person), and the channel/environment. Needless to say, since people obviously communicate in such modes and across such environments quite freely, a superhuman speech recognizer at a minimum must be robust across these dimensions.

30.4.1 Preliminary Recognition Results

In an attempt to get a feeling for the domain sensitivity of current speech recognition systems, a set of experiments was performed. Three corpora were available for training – transaction data [commands, digits, Defense Advanced Research Projects Agency (DARPA) communicator data], switchboard data, and voicemail data. Four corpora were available as test sets – continuous digits, names, switchboard test data (swb98, above) and voicemail test data (vm, above). Systems were trained individually on these corpora. In addition, a common technique – multistyle training – was employed to produce systems in which the training data from the switchboard and voicemail were combined, and one in which all three training corpora were combined. The training and decoding were done as described in Sect. 30.4.2. In all cases, for decoding, a language model suitable for the test data was utilized.

The results are shown in Table 30.8. As can be seen, significant degradation can result when data from one mode of speaking is decoded with models trained on a different mode of speaking. Note also that multistyle training across different corpora to some extent reduces the cross-corpus training degradation effects, but does not always allow one to completely recover the performance levels obtained when test and training data are matched to each other. This implies that blind combination of a huge amount of data from multiple sources is not likely to allow us to reach levels of superhuman performance in cross-domain experiments (and the best way in which to combine language model data has not yet been addressed).

30.4.2 Results on the Multidomain Test Set

A conscious decision was made to focus on telephony transcription data as an initial challenge. Therefore, we continued to utilize the multidomain test corpus described in Sect. 30.2. To deal with the broad range of material present in the multidomain test set, we employed a recognition strategy based on multiple passes of recognition interleaved with unsupervised acoustic model adaptation and on a combination of recognition hypotheses from systems using disparate feature sets and acoustic models. We first describe the basic techniques used in the benchmark system for signal processing, acoustic modeling, adaptation, and language modeling, then we describe the architecture of the recognition system and present the performance of the system on the multidomain test corpus at various stages of processing.

Signal Processing

The systems in this work use either mel-frequency cepstral coefficients (MFCC) or perceptual linear prediction (PLP) features as raw features. The MFCC features are based on a bank of 24 mel filters spanning 0–4.0 kHz. The PLP features are based on a bank of 18 mel filters spanning 0.125–3.8 kHz and use a 12th-order autoregressive analysis to model the auditory spectrum. Both feature sets are based on an initial spectral analysis that uses 25 ms frames smoothed with a Hamming window, a 10 ms frame step, and adds the equivalent of 1 bit of noise to the power spectra as a form of flooring. Both feature sets also are computed using periodogram averaging to reduce the variance of the spectral estimates. The final recognition feature set for all systems are generated by concatenating raw features from nine consecutive frames and projecting to a 60-dimensional (60-D) feature space. The projection is a composition of a discriminant projection (either linear discriminant analysis or heteroscedastic discriminant analysis [30.16]) and a diagonalizing transform [30.17, 18].

Prior to the projection to the final, 60-D recognition feature space, the raw features are normalized. Three

Table 30.8 Preliminary test results. The training corpora are listed across the top and the test corpora are represented by the rows

	Training corpus				
	Transactions	**SWB**	**Voicemail**	**Voicemail + SWB**	**All**
Names	4.4	6.4	8.6		5.3
Digits	1.3	1.9	2.4		1.4
SWB		39	57	46	
Voicemail		47	36	37	

different normalization schemes are used by different systems in this work:

1. utterance-based mean normalization of all features;
2. utterance-based mean normalization of all features except c0 and maximal normalization of c0; and
3. side-based mean and variance normalization of all features except c0 and maximal normalization of c0. In maximal normalization of c0, the maximum value of c0 within an utterance is subtracted from c0 for all frames in the utterance. The estimate of variance is based solely on frames for which c0 exceeds a threshold with respect to the maximum value of c0 in the utterance. This is intended to ensure that the variance is only computed from speech frames.

Acoustic Modeling

We use an alphabet of 45 phones to represent words in the lexicon. Each phone is modeled as a three-state, left-to-right hidden Markov model (HMM). Acoustic variants of the HMM states are identified using decision trees that ask questions about the surrounding phones within an 11-phone context window (±5 phones around the current one). Systems may employ *word-internal* context, in which variants are conditioned only on phones within the current word, or *left* context, in which variants are conditioned on phones within the current and the preceding words.

The majority of the systems described in this work model the leaves of the phonetic decision trees using mixtures of diagonal-covariance Gaussian distributions that are trained using maximum-likelihood estimation (MLE). Subject to a constraint on the maximum number of Gaussians assigned to a leaf, the number of mixture components used to model a leaf is chosen to maximize the Bayesian information criterion (BIC),

$$F(\theta) = \log P(X_s|s,\theta) - \frac{\lambda}{2}|\theta|\log(N_s)\,, \tag{30.1}$$

where $P(X_s|s,\theta)$ is the total likelihood of the data points X_s that align to leaf s under model θ, N_s is the number of such points, and $|\theta|$ is the total number of parameters in model θ. The overall size of an acoustic model may be adjusted by changing the weight on the BIC penalty term, λ. The acoustic models for all recognizers are trained on 247 h of switchboard data and 18 hours of Callhome English data. Early experiments revealed that, after acoustic adaptation, no benefit was obtained by combining other sources of data (such as voicemail data, as described in Sect. 30.4.1) so no additional data was combined with the SWB (switchboard) data.

Two systems described in this work employ alternative acoustic models. One system models leaves using mixtures of diagonal-covariance Gaussian distributions that are discriminatively trained using maximum mutual information estimation (MMIE). In our MMIE training, we collect counts by running the forward–backward algorithm on a statically compiled decoding graph, using beam pruning to constrain the size of the search space [30.19]. This lets us exploit technology developed for fast decoding of conversational speech [30.20] for fast MMIE training as well. The second system models leaves with subspace precision and means (SPAM) models [30.21, 22]. SPAM models provide a framework for interpolating between diagonal-covariance and full-covariance Gaussian mixture models in terms of model complexity and model accuracy. Unlike the diagonal-covariance Gaussian models in this work, SPAM models do not directly use BIC-based model selection.

Canonical Acoustic Models

We use two feature-space transformations, vocal-tract-length normalization (VTLN) [30.23] and maximum-likelihood feature-space regression (FMLLR) [30.24], in an adaptive training framework to train *canonical* acoustic models. The goal of canonical training is to reduce variability in the training data due to speaker- and channel-specific factors, thereby focusing the acoustic model on variability related to linguistic factors. At test time, the feature-space transforms are estimated in an unsupervised fashion, using results from earlier decoding passes.

Our implementation of VTLN uses a set of 21 warp factors that cover a ±20% linear rescaling of the frequency axis. The VTLN frequency warping is applied prior to mel binning in the feature computation. The VTLN warp factor for a speaker is chosen to maximize the likelihood of frames that align to vowels and semivowels under a voicing model that uses a single, full-covariance Gaussian per context-dependent state. Approximate Jacobian compensation of the likelihoods is performed by adding the log determinant of the sum of the outer products of the warped cepstra to the average frame log likelihood.

The FMLLR transformation is an affine transformation of the features in the final, 60-D recognition feature space that maximizes the likelihood of a speaker's data under an acoustic model. FMLLR is equivalent to constrained maximum-likelihood linear regression (MLLR) [30.24], where the MLLR transform is applied to both the means and covariances of the acoustic model. In the remainder of the paper, we will refer to canonical models that use VTLN features as VTLN models and to canonical models that use VTLN features and an FMLLR transformation as SAT (speaker-adaptive trained) models.

Acoustic Model Adaptation

At test time, we also use MLLR adaptation [30.25] of model means to further adapt the recognition system to the specific speaker and environment. Systems that use diagonal Gaussian mixture acoustic models perform two rounds of MLLR. The first round estimates one MLLR transform for all speech models and one MLLR transform for all nonspeech models, and new recognition hypotheses are generated with the adapted models. In the second round, multiple MLLR transforms are estimated using a regression tree and a count threshold of 5000 to create a transform for a regression class. The system using SPAM models performs a single round of adaptation in which a single MLLR transform for all models and a new FMLLR transform are estimated.

Language Modeling and Recognition Lexicon Design

The data used to train the language models consist of 3 million switchboard words, 16 million broadcast news words, 1 million voicemail words and 600,000 call center words. For the initial rescoring of the word internal lattices we used a four-way interpolated language model, each of the components being a back-off 3-gram LM (language model) using modified Kneser–Ney smoothing [30.26]. The mixture weights ($0.45 \cdot \mathrm{Swb} + 0.25 \cdot \mathrm{BN} + 0.15 \cdot \mathrm{VM} + 0.2 \cdot \mathrm{CC}$) are optimized on a held-out set containing 5% of each of the training corpora. For the final rescoring of the left-context lattices the 3-gram mixture components are replaced with 4-gram language models, keeping the mixture weights the same. The 34,000-word vocabulary used in our experiments consists of all the high-count words from our training corpora. The pronunciation dictionary consists of 37,000 entries, yielding a ratio of 1.09 pronunciations per word in the vocabulary. Table 30.9 shows the perplexities and the out-of-vocabulary (OOV) rates for each of the five test sets.

Table 30.9 Perplexities and OOV rates across different test sets

Test set	Perplexity 3gm LM	Perplexity 4gm LM	OOV rate (%)
swb98	94.16	90.08	0.3
mtg	146.45	142.58	0.7
cc1	111.69	106.42	0.3
cc2	52.95	49.30	0.1
vm	94.66	89.32	1.1

Recognition Process and Performance

Recognition of data system proceeded as follows:

P1 Speaker-independent decoding. The system uses mean-normalized MFCC features and an acoustic model consisting of 4078 left context-dependent states and 171 000 mixture components. Decoding is performed using IBM's rank-based stack decoding technology [30.27].

P2 VTLN decoding. VTLN warp factors are estimated for each speaker using forced alignments of the data to the recognition hypotheses from P1, then recognition is performed with a VTLN system that uses mean-normalized PLP features and an acoustic model consisting of 4440 left context-dependent states and 163 000 mixture components. Decoding is performed using IBM's rank-based stack decoder. In the cc2 and vm test sets, which have no speaker information, VTLN warp factors are estimated for individual utterances.

P3 Lattice generation. Initial word lattices are generated with a SAT system that uses mean-normalized PLP features and an acoustic model consisting of 3688 word-internal context-dependent states and 151 000 mixture components. FMLLR transforms are computed using recognition hypotheses from P2. The lattices are generated with a Viterbi decoder. The lattices are then expanded to trigram context, rescored with a trigram language model and pruned. In the cc2 and vm test sets, which have no speaker information, FMLLR transforms are estimated for individual utterances.

P4 Acoustic rescoring with large SAT models. The lattices from P3 are rescored with five different SAT acoustic models and pruned. The acoustic models are as follows:

- [A] An MMIE PLP system consisting of 10 437 left context-dependent states and 623 000 mixture components. This system uses maximum-normalization of c0 and side-based mean and variance normalization of all other raw features.
- [B] An MLE PLP system identical to the system of P4A, except for the use of MLE training of the acoustic model.
- [C] An MLE PLP system consisting of 10 450 left context-dependent states and 589 000 mixture components. This system uses mean normalization of all raw features.

Table 30.10 Word error rates (percentage) for the components of the multidomain test set and the overall, average error rate for the corpus. For passes where multiple systems are used (P4–6), the best error rate for a test component is highlighted

Pass	swb98	mtg	cc1	cc2	vm	All
P1	42.5	62.2	67.8	47.6	35.4	51.1
P2	38.7	53.7	56.9	44.1	31.7	45.0
P3	36.0	44.6	46.6	40.1	28.0	39.1
P4A	**31.5**	**39.4**	41.7	38.2	26.7	35.5
P4B	32.3	40.0	**41.3**	39.0	26.7	35.9
P4C	32.5	40.2	42.1	39.9	27.0	36.3
P4D	31.7	40.3	42.6	**37.6**	**25.8**	35.6
P4E	33.0	40.5	43.4	38.8	26.9	36.5
P5A	30.9	**38.3**	**39.4**	36.9	26.1	34.3
P5B	31.5	38.5	**39.4**	37.0	26.5	34.6
P5C	31.6	38.7	41.0	39.4	26.8	35.5
P5D	**30.8**	39.0	41.1	**36.7**	**25.6**	34.6
P5E	32.1	38.9	41.8	36.8	26.4	35.2
P6A	**30.4**	**38.0**	**38.9**	36.5	25.7	33.9
P6B	31.0	38.3	**38.9**	36.4	25.8	34.1
P6C	31.2	38.4	40.1	38.9	26.3	35.0
P6D	**30.4**	38.6	40.8	36.3	**25.5**	34.3
P6E	31.5	38.5	41.6	**35.9**	25.7	34.6
P7	29.0	35.0	37.9	33.6	24.5	32.0

- [D] A SPAM MFCC system consisting of 10 133 left context-dependent states and 217 000 mixture components. The SPAM models use a 120-dimensional basis for the precision matrices. This system uses mean normalization of all raw features.
- [E] An MLE MFCC system consisting of 10 441 left context-dependent states and 600 000 mixture components. This system uses maximum-normalization of c0 and mean normalization of all other raw features.

  The FMLLR transforms for each of the five acoustic models are computed from the one-best hypotheses in the lattices from P3. FMLLR transforms are estimated for individual utterances in the vm test set, but on the cc2 test set a single FMLLR transform is estimated from all utterances. The vm test set contains many long utterances [30.28], and the FMLLR estimation procedure has sufficient data, even with very large acoustic models. We found that the cc2 test set contained only very short utterances, and the FMLLR procedure failed to converge on many utterances with the large acoustic models.

P5 Acoustic model adaptation. Each of the five acoustic models are adapted using one-best hypotheses from their respective lattices generated in P4; no cross-system adaptation is performed. As described above, the systems using Gaussian mixture acoustic models are adapted using two sets of MLLR transforms, while the SPAM acoustic model is adapted using an FMLLR transform and an MLLR transform. The lattices from P3 are rescored using the adapted acoustic models and pruned. As in P4, transforms are estimated for individual utterances in the vm test set, but are estimated globally for the cc2 test set.

P6 4-gram rescoring. Each of the five sets of lattices from P5 are rescored and pruned using a 4-gram language model.

P7 Confusion network combination. Each of the five sets of lattices from P6 are processed to generate confusion networks [30.29], then a final recognition hypothesis is generated by combining the confusion networks for each utterance.

The performance of the various recognition passes on the test set is summarized in Table 30.10.

Conclusions

Reasonable recognition performance can be obtained on a broad sample of conversational American English tasks using acoustic models trained only on switchboard and Callhome data. The results on the mtg set illustrate this point most strongly, for neither the acoustic models nor the language models are trained on meeting data. This supports the observation that the switchboard corpus is representative of the acoustic–phonetic and stylistic properties of conversational American English [30.30].

Multipass decoding with unsupervised adaptation and a combination of disparate systems are effective techniques for achieving good recognition performance on diverse data sources. On this test set, they can reduce the overall error rate from 51.1 to 32.0%.

While system combination can provide consistent gains in recognition performance, they are relatively small for the rather substantial amount of computation incurred. Had we used only the MMIE PLP system and performed consensus decoding instead of the confusion network combination in P7, the overall error rate on the test would have increased to 33.1%. This rather intensive amount of computation is a clear impediment to research and resulted in a major redesign of the system as described in the next section.

30.4.3 System Redesign

As described above, the main problem with the above system was its high level of complexity. Too many passes over too many systems are required – the above system ran in about 5000×RT (real-time), which is really outrageous. If one looks at the culprits in the computation, it is primarily the need to generate many sequences of lattices inefficiently because of an inability to handle all the recognition sources of knowledge in a single pass. For this reason a major redesign of the recognition system was performed to allow for all the knowledge sources needed for recognition to be compiled into a single static network, and many of the multiple passes used to determine warping factors and FMLLR transforms were combined into a single pass or eliminated altogether [30.20]. It was found that this combination not only sped up decoding enormously but also resulted in much better accuracy. Essentially, each pass of lattice generation and pruning was introducing search errors that were eliminated by migrating to a single-pass decoding process. Another innovation that was added was incorporation of a dynamic language model that could automatically modify the interpolation weights in a two-pass recognition process. Each utterance was decoded with a general language model and a lattice was produced. The interpolation weights of the LM components were then adjusted to minimize perplexity of the decoded utterance text, and the new LM that resulted was used to rescore the lattice to produce the final recognition result.

Table 30.11 Recognition results on the multidomain test set after system redesign

Corpus	Previous	After redesign
swb	29.0	27.0
mtg	35.0	33.6
cc1	33.6	24.7
cc2	37.9	36.9
vm	24.5	20.9
Average	32.0	28.6
Number of recognizers	5	1
Speed	5000×RT	20×RT

The results are shown in Table 30.11. As can be seen, not only has the recognition accuracy significantly improved, but the amount of computation dramatically has also decreased. The need for multiple recognizers is essentially eliminated and even the speed of a single recognition pass significantly dropped. Analysis indicated that approximately 2% absolute of the performance gain was attributable to the presence of the adaptive language model.

30.4.4 Coda

After the above set of experiments were completed, a shift in emphasis to focus on public data, such as those present in the DARPA EARS (effective, affordable, reusable speech-to-text) and GALE (global autonomous language exploitation) programs, took place. The 2004 EARS system [30.14] was basically an extension of the system described in Sect. 30.4.3 to three systems (speaker independent, speaker adaptive with diagonal covariances, and speaker adaptive with full covariances), all trained with MPE (minimum phone error) [30.31] and fMPE [30.32]. In addition, significantly more data (2000 h) were used to train the models. Unfortunately, as described in Sect. 30.4.3 the signal processing needed to process the superhuman test set was a modified version of the EARS system signal processing to improve robustness, so it was only possible to decode the swb data component with the EARS system. The result was a 19% word error rate compared to a 27% word error rate on the swb component of the system described in Sect. 30.4.3. Approximately half of the improvement could be attributed to increased data and the rest divided across the new acoustic modeling (MPE and fMPE) and the use of multiple systems. Assuming that at least some of these improvements would have held up across the other data components, it is safe to say we would have been looking at error rates between 20% and 25% on the overall superhuman corpus. Given the results of the human listening tests (Sect. 30.3.2), at least on short segments of speech, one could almost say that for telephony speech transcription and limited acoustic context, superhuman performance no longer seems completely out of reach.

30.5 Speculation

Although the above sections suggest some progress has been made in the direction of achieving superhuman speech recognition, it is clear that there is still a long path ahead. First, system performance even on a subset of domains (telephony transcription) – say at best 20% WER (word error rate) – is still a good factor of three or four away from human performance on this task – best reported results being about 5%. Second, except for the preliminary results described in Sect. 30.4.1, no serious attempts have yet been made to develop a single system simultaneously capable of handling transactional and transcription data. Lastly, the robustness of today's speech systems with respect to varying channels and noise is still weak. Is it possible, then, to achieve superhuman speech recognition by plugging away at the essentially incremental approaches that have typically resulted in incremental gains on the order of 10–20% a year that have been described in the rest of the paper, with no major insights from human performance?

Today's speech recognition systems have all evolved to have the same basic structure: spectral features are periodically sampled at the frame level from the speech signal, and the probability distribution of these feature vectors is captured by hidden Markov models (HMMs). The HMMs are combined with a language model of the short-term dependencies of word sequences, and the word hypotheses are produced by various efficient dynamic-programming-based search mechanisms [30.33]. Many alternatives and enhancements have been tried over the years: articulatory models, neural networks, segment models, trigger language models, structured language models, but all have either failed or not shown significant advantages over this basic structure. In addition, regular significant improvements have always resulted over the years via refinements to the basic structure, placing a successively increasing barrier to entry on alternative formulations.

There is no question that advances in speech recognition will still continue via incremental techniques – one would be foolish in the face of essentially 30 years of incremental improvements to deny that progress has not been continually made. However, this does not rule out exploration of alternative methodologies in parallel or on top of existing high-performance recognition systems. In the remainder of this chapter, some speculation is presented both about some promising incremental approaches with good payoffs, as well as some more-dramatic changes in speech recognition methodologies.

30.5.1 Proposed Human Listening Experiments

The listening experiments described above only scratched the surface in terms of trying to tease out what the most useful research directions in speech recognition might be. An extension of the above described experiments is the following.

1. Word sequences are generated automatically from a trigram LM, and then read aloud.
2. The sentences are decoded with an ASR (automatic speech recognition) system, and the WER is computed.
3. Noise is added to the to the recordings until human listeners have the same WER, i. e., the acoustics are calibrated to equalize human and machine performance.
4. Meaningful sentences are recorded, read, and decoded by the ASR system.
5. The same level of noise is added to the same meaningful sentences, which are then transcribed by human listeners.

When meaning is added to the sentences, there will be a differential impact on human and machine performance, presumably with humans benefiting more. This difference will provide a meaningful measure of the amount that can be gained by moving beyond trigram LMs and using broad syntactic, semantic, and pragmatic knowledge sources.

Yet another problem is that there is a mismatch between the metrics that are used to measure ASR performance, and the intended use of these systems. Performance is measured with word error rate, which is dominated by errors in function words. For example, the five most common errors in a typical large-vocabulary recognition system are:

1. and/in
2. it/that
3. was/is
4. that/the
5. the/a

While these words are very common, it is not clear that recognizing them is always critical to the performance of a real application. For example, in a typical call-center application, a user might call to make a travel reservation and say something like *I want to fly to Boston on Tuesday.* What the system needs to produce is a representation something like (what = reservation; to = Boston; date = Tuesday; from = ????) and realize that an appropriate response is something like *Where are you flying from?* It does not matter, however, if there are minor recognition errors like *I want to fly to Boston at Tuesday,* or even more serious ones like *I went to fly to Boston on Tuesday,* as long as the system is able to recover the underlying meaning with confidence. In order to understand the relationship between word error and concept error, we propose to conduct a final set of experiments to flesh out this relationship.

One such experiment involves revisiting the confusion network experiments described earlier, but in the context of question-answering. In this experiment, subjects are shown confusion networks and then asked to answer questions. If they are able to do this on the basis of the sausages, then we are warranted in taking confusion networks – without further acoustic processing – as the basis for high-level linguistic processing, regardless of WER.

A second experiment revisits the notion of *calibrated acoustics*, but again in the context of question-answering. Here, increasing noise is added to utterances, and humans are asked to both transcribe the utterances and answer questions. Presumably, people will be able to answer questions even after the WER of their transcripts has degraded considerably, indicating that WER is not a good measure of the information transmitted, and suggesting the use of concept error or slot-filling error as a better metric.

In summary, by conducting a series of human listening and comprehension experiments, one can gain a better understanding of how to improve our current systems, what the fundamental limitations of different knowledge sources are, and how to measure system performance.

In addition to human listening experiments, it might be possible to establish other fundamental bounds, e.g., of orders of magnitude extensions of data sets for current basic structure language models. For instance, it is possible to use a Google query interface [30.34] to check the frequency of word sequences appearing in the lattice or *sausage*, i. e., to see if and how often corresponding 5-grams, 4-grams, 3-grams and 2-grams appear in the Google index. Such experiments would help to answer questions about the value of having four orders of magnitude more data for language modeling.

30.5.2 Promising Incremental Approaches

Modeling Spontaneous Speech

The main problem with spontaneous speech is that it contains substantial segmental and suprasegmental variations not present in read speech (which, of course, is much easier to collect in bulk), such as pitch and duration variations, hesitations, false starts, ungrammatical constructs, emotionally expressive speech (laughter, etc.). The high levels of human performance that can be obtained from just auditory presentation of speech out of context [30.4] suggest that the only clues that are needed to decipher speech lie in the local speech signal. In extremely noisy environments, other cues, such as visual ones, may be required. Consequently, if the WER performance is bad, it must be the case that

1. the stochastic models used to model the observation are not accurate;
2. current feature-extraction techniques (mel cepstra) do not extract all the necessary information;
3. the language model is inadequate; or most likely,
4. all of the above.

Most of the speech recognition systems today use frame-level observations. Though some work has been done on suprasegmental feature extraction [30.35], it has enjoyed only a limited amount of success. One possible reason is because neither frame-level observations nor suprasegmental observations are by themselves sufficient to do an adequate job of modeling (spontaneous) speech. Segmental models [30.36, 37] make fewer assumptions than HMMs and provide a better modeling framework by modeling sequences of observations rather than individual observations; however, the goal of multiscale modeling requires a framework that is more powerful than segmental models. The end objective of any model is to compute the joint probability of all observations. This joint probability can be expressed as a product of several conditional probabilities

– however our goal is to make this factorization different for different observations. Graphical models provide a mechanism whereby different factorizations of a joint distribution can be specified by means of a directed graph [30.38].

To model the statistical dependencies in the extracted features a graphical model that specifies the factorization of a joint distribution by means of a graph can be used. This framework can be used to model arbitrary dependencies, including temporal dependencies in different observation streams. Appropriate dependencies can be selected via correlation or mutual information techniques [30.39, 40]. In addition to these statistically identified dependencies, linguistic knowledge related to the speech production process can also easily be incorporated into the graphical model – for instance, the formant frequency estimates at a given time frame could provide valuable information about the identity of the phone at the current and adjacent time frames. Also, constraints on the formant trajectories between adjacent time frames (for continuity reasons) can easily be incorporated into graphical models.

Such a scheme was recently tried [30.41] and demonstrated improvements on the switchboard corpus. A particularly attractive aspect of such a model is that it can easily be augmented with additional suprasegmental features, such as say pitch, which might prove valuable for modeling syllables and words in tonal languages such as Mandarin, or for better spotting of disfluencies.

Multi-Environment Systems

Another objective in achieving human performance is to develop a system that will work on several different types of speech: full bandwidth, telephone bandwidth, or recorded with close-talking or far-field microphones. One possibility is to bandlimit all sources of speech data and train a system using this data, however, this would provide inferior performance for cases where full-bandwidth or low-noise speech is available. Another possibility is just to run multiple models in parallel with a switch that detects bandwidth changes, as is done in many of today's systems that process broadcast news, but that approach completely fragments data into small pieces, generating a complex multicomponent system that is very hard to manage. Consequently, the goal is to either: (i) develop a single modeling framework that can make use of all the information in full-bandwidth speech when it is available, and can also deal with bandlimited speech, or (ii) investigate features that are insensitive to bandwidth-related distortions.

A possible modeling framework for (i) is to partition the observation vector into two components, one of which is always available, $o_{1,t}$, while the other may be hidden, $o_{2,t}$. The formulation is based on the fact that, if the global joint statistics of o_1 and o_2 are known, and for each class the probability density of o_1 and o_2 are known, then it is possible to predict the missing observation from the joint statistics.

The basic speech recognition problem is to find the joint probability of the observation sequence $o_{1,1} \cdots o_{1,T}$, and, when it is available, $o_{2,1} \cdots o_{2,T}$. As the observation probability is conditioned on a state sequence s_1^T that is hidden, this may be written as

$$\begin{aligned} p\left(o_{1,1}^T o_{2,1}^T\right) &= \sum_{s_1^T} p\left(o_{1,1}^T o_{2,1}^T, s_1^T\right) \\ &= \sum_{s_1^T} p\left(o_{2,1}^T / o_{1,1}^T, s_1^T\right) p\left(o_{1,1}^T / s_1^T\right) p\left(s_1^T\right) \end{aligned} \tag{30.2}$$

If $o_{2,1}^T$ is observable, we can approximate the term $p(o_{2,1}^T / o_{1,1}^T, s_1^T)$ by $p(o_{2,1}^T / s_1^T)$, and if $o_{2,1}^T$ is not observable, we can approximate it by $\hat{o}_{2,1}^T$, where $\hat{o}_{2,1}^T = p(o_{2,1}^T / o_{1,1}^T)$, and the joint statistics of $o_{1,1}^T, o_{2,1}^T$ are used to compute the latter term. (For a Gaussian joint distribution, $\hat{o}_2$ turns out to be a linear operation on o_1.) Hence, the global joint statistics of o_1, o_2 are used to predict the value of o_2, making it possible to use information related to the probability distribution function (PDF) of o_2 for the class s, even when o_2 is not observed.

A generalization of this approach was successfully used in [30.42] to improve robustness for speech in very high-noise environments. In such noise, large sections of the time–frequency display for a speech signal are often masked by the noise, not just the higher frequencies as in telephony speech. It was shown that variations on the above technique could improve the effective signal-to-noise (S/N) ratio by more than 10 dB in certain cases. However, the technique has not yet been applied to bandlimited speech.

Possible frameworks for (ii) include the following. [30.40] investigated the utility of a *spectral-peak* feature that is closely related to the formant frequencies, and initial experimental results indicate that the spectral peak features definitely carry information that can help speech recognition even in clean conditions. These initial experiments were based on simple-minded *feature fusion* of the spectral peak features with the cepstra, and going beyond this simple approach to the more-sophisticated multiscale graphical models described above, which en-able trajectories to be modeled in the formant space, will lead to powerful and robust models for speech recognition.

Another approach to (ii) is to adapt the models or features to the new environmental conditions. In addition to popular techniques such as MLLR and FMLLR, and more recently, fMPE, another approach is to use a nonparametric mapping approach for the cumulative probability distribution function [30.43]. Assume that it is possible to transform the adaptation features $x' = g(x)$ in such a way that the cumulative distribution function (CDF) of the transformed adaptation data is made identical to the CDF of the training data. This criterion is used to guide the design of the transformation. Falling as it does into the category of *data transformation* techniques, this method is independent of the modeling framework that is used; consequently, it can be easily incorporated into the segmental graphical model framework suggested earlier. In [30.44], this method was applied to the speech samples for the application of speaker identification; here this method is extended to deal with multidimensional observations that are used in speech recognition.

This approach is motivated by the following reasoning. It is well known that the assumed parameterization of the true PDF of the data is often inaccurate. Consequently, if we could deal with the empirically observed CDF of the data, there would be no need to make any modeling assumptions. Secondly, the CDF is a well-behaved function (monotonic and lying between 0 and 1), consequently it is relatively easy to define a mapping of the feature dimension that equates two CDFs. Thirdly, this method is computationally much simpler than most other adaptation schemes. Most adaptation schemes require the adaptation data from the test environment to be transcribed, with an associated overhead. The CDF matching scheme on the other hand does not require any prior transcription, and the process of computing the nonlinear transformation is also relatively inexpensive.

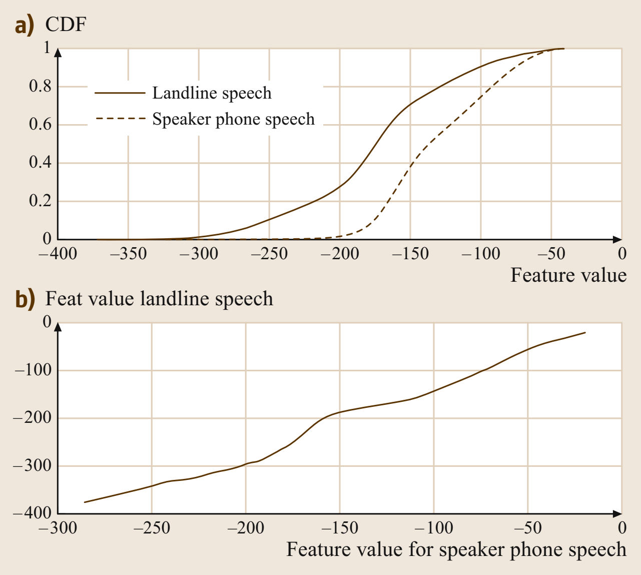


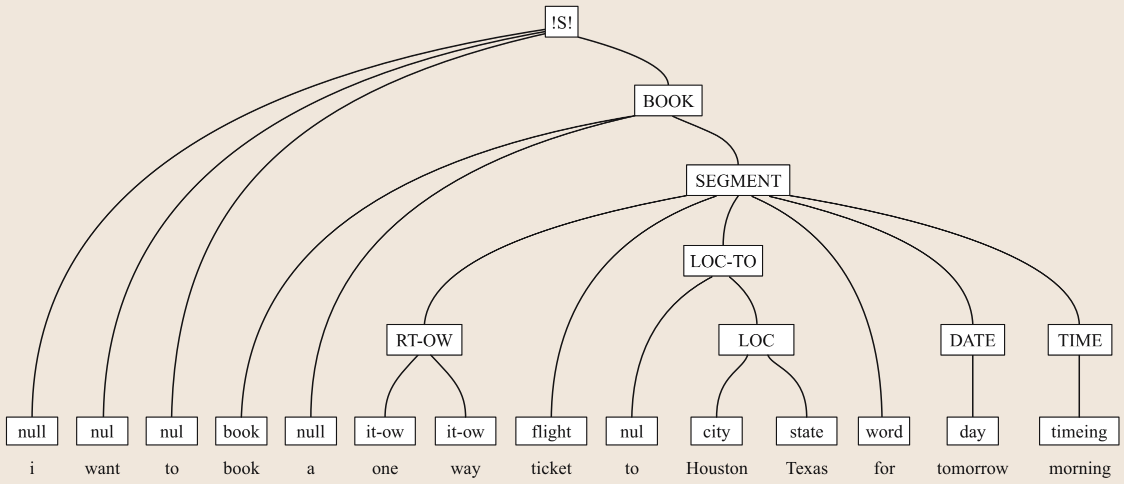
Fig. 30.4a,b CDF matching for landline and speakerphone features. **(a)** CDF functions for landline and speakerphone data, **(b)**the input–output mapping

For the case of multidimensional features, if the dimensions are independent, then the CDF-matching technique can be applied to each dimension independently. However, as this is not generally the case, it is necessary first to transform the feature into a space where the dimensions are uncorrelated. This needs to be done simultaneously for both the training and adaptation data. The solution to this problem is a transformation consisting of the generalized eigenvectors of the covariance matrices of the training and test data.

This technique has been been used to compensate for the mismatch between landline and speakerphone telephone data. In Fig. 30.4, the CDF for the first cepstral dimensions for landline and speakerphone data is shown, with the corresponding mapping. In preliminary experiments, the CDF mapping technique improved performance on speakerphone data by 30% relative [30.43].

Improved Language Modeling

In the last few years, a number of research results by IBM [30.45, 46] and other research groups [30.47–49] have appeared that strongly suggest that linguistic information, such as meaning and language structure, can be used to improve speech recognition performance significantly. The scale of these research activities and experiments is limited and recognition experiments have tended to be performed only on narrow domains. However, the improvements are encouraging. For example, the semantic structured language models [30.45, 46] shown in Fig. 30.5 have reduced WER by 20%, and increased confidence by 30–60% for multiple ASR task domains, such as air travel, finance, medical and military domains, compared to when no semantic structure information was used. In such a language model, the N-best hypotheses are first produced by the recognizer. For each hypothesis, the recognized words are then mapped onto a set of *classes*. Then, a semantic parse of the classed words in each hypothesis is produced, and a maximum-entropy model is constructed in which the word w_j is



- [!S! i want to [BOOK bock a [SEGMENT [RT-OW one way RT-OW] ticket [LOC-TO to [LOC Houston Texas LOC] LOC-TO] for [DATE tomorrow DATE] [TIME morning TIME] SEGMENT] BOOK] !S!]
- 6 history question types: (default), ($w_{j\text{-}1}$), ($w_{j\text{-}1}$, $w_{j\text{-}2}$), (L,N), (O,M), (L,G)
- Contrast to N-Gram LMs:

  + i want to book a one way ticket to Houston Texas for tomorrow morning

Fig. 30.5 A semantic structured language model. The words are shown at the *bottom* of the parse tree. The next level up represents the output of the word classer, and the higher levels represent the parse tree constituents. At the bottom is shown a text-based bracketed representation useful for determining language model dependencies (see text)

predicted as

$$p\left(w_j|W_1^{j-1}\right) = p(w_j|w_{j-1}, w_{j-2}, p_j, g_j, c_j) , \tag{30.3}$$

where p_j is the parent constituent of the current word, c_j is the last complete constituent, and g_j is the constituent identity of the highest-level concept below the sentence level in the semantic parse tree. For example, in Fig. 30.5 the parent constituent of *tomorrow* is *DATE*, the last complete constituent for *morning* is *DATE*, and the highest level concept for *morning* is *BOOK* (note that the lower boxes are the class assignments and not the parse tree constituents). The resultant language model is then used to rescore the N-best hypotheses.

In these experiments, only a small number of linguistic features, among the almost unlimited number of them, have been investigated for speech recognition. Typically, they are restricted to extracting syntactic or semantic attributes, and fitting them into the basic structure of speech recognition, i. e., adding semantic or structural attributes into a trigram or bigram language model and computing the new perplexity or measuring the resulting WER. These research activities are mainly in the area of implementation or engineering, rather than in developing new computational models that can naturally marry automatic speech recognition and computational linguistics. Only the latter can be expected to result in dramatic ASR performance improvements.

Such efforts should come naturally, as the types of information, i. e., acoustic (mainly statistical) and linguistic (mainly structural), are complementary. Even though creating complete new models for speech recognition is difficult, there are examples of leveraging the combined power of statistical and structural information – namely, in the field of natural language processing. Taking insights from such areas of computational linguistics such as word sense disambiguation, text meaning representation, and syntactic parsing, and combining them with the power of statistical techniques, new models have been created and successfully applied in text analytics, search and classification. Often these models use new knowledge resources, such as the Penn TreeBank, the Brown Corpus, WordNet, and the extended WordNet.

Based on these observations, two parallel lines of research are suggested in order to incorporate semantic and syntactic knowledge into ASR:

1. Investigate the use of a larger number of linguistic features, such as lexical features, semantic and syntactic relation and structural features, morphological features, part–whole relationship features, and experiment with more task domains of ASR. Gradually extend to broad domains of ASR. These will test the broad validity of the claims about the importance of semantic and structural information as well as help to find the set of useful features.
2. Investigate new computational models for speech recognition in which natural language structures and linguistic features are naturally integrated in one unified framework. The direct model proposed by IBM [30.50] is one possible approach. In this framework, a model is constructed in which the probability of a phonetic state, word, or sentence is computed as a direct function of the underlying knowledge sources including both linguistic and acoustic information. Figure 30.6 depicts such a model in which the probability of the next state of the model is a function of the word, the sentence, and the acoustic observation. This is contrasted with current practice, which uses a generative model to compute the probability of a set of features from an underlying word or sentence in conjunction with an independent language model. The new models could be used as one of the specialists in the doctor/specialist framework described below.

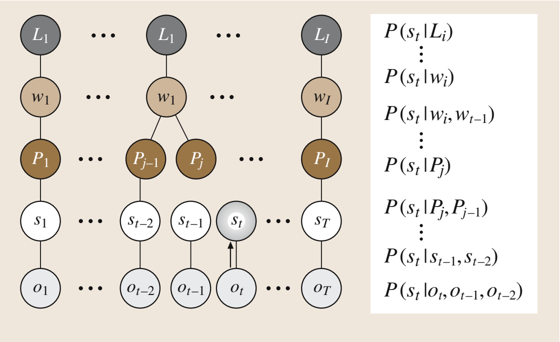


Fig. 30.6 A direct model for automatic speech recognition. Here, the probability of the state is directly conditioned on the associated phoneme, the word, the language, and the observations

30.5.3 Promising Disruptive Approaches

The Doctor/Specialists Paradigm

State-of-the-art ASR systems adopt a monolithic approach to acoustic and linguistic knowledge sources. For example, the same acoustic features and model are used to distinguish all acoustic confusions [30.33]. It is quite remarkable that this approach to ASR, chosen for its simplicity and lack of a better alternative, works so well in practice. However, it is also clear that for distinguishing particular acoustically confusable units, e.g., *a* versus *the*, a specialist classifier trained only to distinguish between the confusable units will perform significantly better. One solution is the design and development of a *doctor/specialists* framework for ASR. The basic idea is as follows. The doctor, who is responsible for the overall recognition process, will partition the speech into regions of certainty and regions of confusion, which can be done with the current state of the art. In regions of confusion, inputs from specialists (acoustic, linguistic, and visual) who are experts at distinguishing between particular sets of words or phonemes will be invoked and integrated in a systematic manner.

A key ingredient in ASR improvements over the past decade has been the inclusion of increasing amounts of contextual information – both acoustic and linguistic, and the framework above can be viewed as a significant step towards exploiting larger contextual information. The doctor uses all the contextual information available at a given point to decide on what the next local problem to be resolved is (e.g., is it *f* as in *fine* or *sh* as in *shine*) and then consults the best specialist for this problem. This approach to ASR may be significantly better than the current approach for the following reasons. First, specialists can use more-complex and discriminative models than those in use today [30.51, 52]. Specifically, for the specialists one can replace the generative (HMM-based) models in use today with the discriminative models that have been hard to use in ASR because of the large number of acoustic classes being modeled simultaneously. Secondly, specialists need only be trained in regions difficult for the doctor. Thirdly, specialists can focus on novel acoustic, linguistic and visual (when available) features specific to the confusions they have to resolve. Finally, because of the smaller size of the problems they tackle, specialists can use computationally and memory intensive template-matching approaches for resolving confusions. This methodology may also entail using orders of magnitude more acoustic training data than used by current systems (e.g., 100 000 h of speech). Some initial promising results on such a technique have been obtained in [30.53], where the term *acoustic codebreaking* is used to describe the doctor/specialists procedure.

In principle this doctor/specialists approach outlined requires a completely new search strategy (doctor) and associated modeling strategies (specialists). In practice, we believe this strategy is well suited for implementation in the context of the current state of the art. Current ASR systems can generate confidence-weighted word lattices or sequences of word confusion sets (or sausages, see Fig. 30.1) [30.29]. The oracle word error rate on lattices and/or sausages can be significantly lower than the one-best word error rate, e.g., 10% versus 30%. This implies that resolving confusions in lattices or sausages will result in a dramatic gain in accuracy. To exploit this one would require

- research on the construction of acoustic, linguistic, and visual specialists to distinguish between phonemes and/or words that typically co-occur in sausages
- research on when to invoke and how to integrate input from the specialists

This manifestation of this paradigm naturally decouples the recognition process into two tasks, each of which can be the thrust of independent research programs:

- the generation of minimal size lattices and/or sausages with zero oracle word error rate, and
- the choice of the one-best word hypothesis in lattices and/or sausages with the help of specialists adapted to the particular confusions present in a given lattice or sausage.

Large-Scale Information Fusion

The current state of the art in speech analysis operates in a highly stylized way: the waveform is processed in 25 millisecond chunks – one every 10 milliseconds – regardless of the acoustic conditions, and linguistic knowledge is captured at the level of word N-grams only. It is known, however, that in animal auditory processing there are hundreds of cells and neural regions that are specially adapted to respond to specific acoustic and linguistic stimuli [30.54], and that somehow these features are combined on an extremely large scale to produce speech perception. Drawing loosely on this analogy, another disruptive approach to speech recognition is to develop methods for extracting and combining extremely large numbers of complimentary acoustic and linguistic features. Figure 30.7 illustrates some of the acoustic aspects of this paradigm.

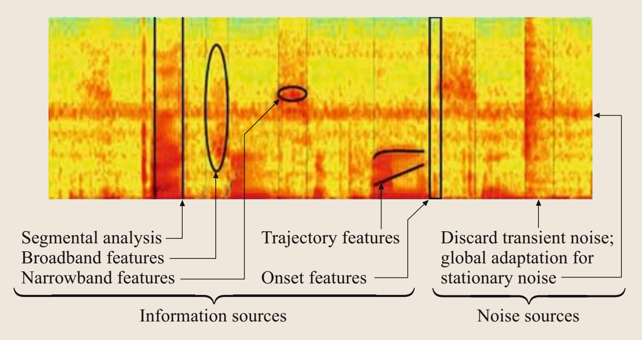


Fig. 30.7 Multiscale and variable-rate features

Here, the spectrum corresponding to a word has been segmented into variable-length components, both vertically and horizontally. The first oval represents a broadband feature measuring the energy across a wide range of frequencies. The second oval, in contrast, is a narrowband feature, while the dark rectangle at the right measures the rate of change of energy across all frequencies, resulting in an onset detector. The individual curved lines track the formants of one syllable, and their temporal evolution results in two trajectory features. The long horizontal band of orange in the middle of the figure is hypothesized to be noise, and adaptation is used to remove its effects. Once a large number of features has been defined and evaluated, it still remains to combine them in a coherent probabilistic model, and here the maximum-entropy approach is ideally suited. The values of arbitrary numbers of both acoustic and linguistic features can be combined in a theoretically sound fashion to produce posterior estimates of word probabilities.

Although conceived as using far more features, and combining them in a less-supervised way, this model is closely related to the doctor/specialists model. The integrating probabilistic model, e.g., maximum entropy, acts essentially as a word-level doctor and combines features to determine word identity on a word-by-word basis. The feature values represent specialist outputs, and maximum entropy provides a framework for weighting them.

Large-Scale Decoder Combination

Recent competition-grade ASR systems have begun to combine the outputs of several systems in a voting process [30.55]. In this strategy, a system actually consists of four or five independently constructed ASR system. Each of these is handcrafted to differ from the others in some minor details, for example the use of MFCC as opposed to PLP features, or the set of phonetic units that is used. Typically these systems make somewhat uncorrelated errors, and therefore a voting strategy tends to converge on the correct answer. Often, the systems represent a failed attempt at building a better single decoder that results simply in something that makes equally many but different errors.

This basic strategy has been proven to create gains, but ignores one of the major developments in machine learning – the idea, termed boosting, that hundreds or thousands of systems can actually be manufactured and combined in a fully automated way so as to create mathematically provable bounds on the error rate of the resulting composite [30.56]. In this framework, a sequence of classifiers is constructed such that at each iteration the latest classifier tends to correct the errors of the previous ones. This is done by maintaining a carefully selected weight on each of the training examples so that it is possible to prove that, as long as a better-than-chance classifier can be built at each iteration, the error rate will drop to zero in an exponentially decreasing fashion. On the surface this may appear to generate a system of extreme complexity: a 40 dB complexity increase relative to the system described in Sect. 30.4.2. The key challenge is to develop a set of systems automatically that all have a similar structure whose overall management is similar. An initial start in this direction was described in [30.57], in which a set of similar systems were generated from the same data by randomizing the data used to initialize the state context acoustic decision trees. Significant recognition improvements were demonstrated on the switchboard task by combining up to six similar randomized systems. This can be generalized to other components of the system – say the signal processing and language models – to generate many parallel systems. Of course, boosting is not a panacea and a major issue will be how to generate appropriate sets of complementary systems to be combined.

In order to achieve such goals, the computing power of large clusters of workstations – using new architectures such as Blue Gene [30.58] and Cell [30.59] – can be used to build and combine decoders on a scale that has not been conceived before. Doing this will require full parameterization of the process of building a decoder (within the basic structure) so that system builds that utilize different subsets of the training data will be able to span the necessary space of acoustic and language models. Perhaps in the end this is the most promising approach of all, insofar as it begins to address building systems that even approach the complexity and processing power of the human brain.

References

30.1 J.G. Fiscus, W.M. Fisher, A.F. Martin, M.A. Przybocki, D.S. Pallett: 2000 NIST evaluation of conversational speech recognition over the telephone, Proc. 2000 Speech Transcription Workshop (2000)

30.2 A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters: The ICSI Meeting corpus, Proc. ICASSP, Vol. I (2003) pp. 364–367

30.3 M. Padmanabhan, G. Saon, J. Huang, B. Kingsbury, L. Mangu: Automatic speech recognition performance on a voicemail transcription task, IEEE Trans. Speech Audio Process. **10**(7), 433–442(2002)

30.4 R.P. Lippmann: Speech recognition by machines and humans, Speech Commun. **22**(1), 1–15 (1997)

30.5 I. Pollack, J.M. Pickett: The intelligibility of excerpts from conversation, Lang. Speech **6**, 165–171 (1963)

30.6 E. Chang, R. Lippmann: Improving wordspotting performance with artificially generated data, Proc. ICASSP, Vol. 1 (1996) pp. 526–529

30.7 J.B. Allen: How do humans process and recognize speech?, IEEE Trans. Speech Audio Process. **2**(4), 567–577 (1994)

30.8 C.E. Shannon: Prediction and entropy of printed English, Bell Syst. Tech. J. **30**, 50–64 (1950)

30.9 NIST Speech Group: The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan, Version 4 (2003) http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf

30.10 W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajič, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, W.-J. Zhu: Automatic recognition of spontaneous speech for access to multilingual oral history archives, IEEE Trans. Speech Audio Process. **12**(4), 420–435 (2004)

30.11 I. Nastajus: http://en.wikipedia.org/wiki/Naturally Speeking (2007)

30.12 P. Woodland, H.Y. Chan, G. Evermann, M.J.F. Gales, D.Y. Kim, X.A. Liu, D. Mrva, K.C. Sim, L. Wang, K. Yu, J. Makhoul, R. Schwartz, L. Nguyen, S. Matsoukas, B. Xiang, M. Afify, S. Abdou, J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, F. Lefevre, D. Vergyri, W. Wang, J. Zheng, A. Venkataraman, R.R. Gadde, A. Stolcke: SuperEARS: Multi-Site Broadcast News System, DARPA EARS 2004 Workshop (2007), http://www.sainc.com/richtrans2004/uploads/monday/EARS BN Super team.pdf

30.13 A. Aaron, S. Chen, P. Cohen, S. Dharanipragada, E. Eide, M. Franz, J.-M. Leroux, X. Luo, B. Maison, L. Mangu, T. Mathes, M. Novak, P. Olsen, M. Picheny, H. Printz, B. Ramabhadran, A. Sakrajda, G. Saon, B. Tydlitat, K. Visweswariah, D. Yuk: Speech recognition for DARPA Communicator, Proc. ICASSP, Vol. 1 (2001) pp. 489–492

30.14 H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, G. Zweig: The IBM 2004 conversational telephony system for rich transcription, Proc. ICASSP, Vol. 1 (2005) pp. 205–208

30.15 J. Fiscus: The Rich Transcription Spring 2006 (RT-06S) Evaluation Results (NIST Speech Group, 2007) http://www.nist.gov/speech/tests/rt/rt2006/spring/pdfs/rt06s-STT-results-v7.pdf

30.16 G. Saon, M. Padmanbhan, R. Gopinath, S. Chen: Maximum likelihood discriminant feature spaces, Proc. ICASSP, Vol. II (2000) pp. 1129–1132

30.17 R.A. Gopinath: Maximum likelihood modeling with Gaussian distributions for classification, Proc. ICASSP, Vol. 2 (1998) pp. 661–664

30.18 M.J.F. Gales: *Semi-tied full-covariance matrices for hidden Markov models*, Vol. CUED/F-INFENG/TR287 (Cambridge Univ. Engineering Department, Cambridge 1997)

30.19 J. Huang, B. Kingsbury, L. Mangu, G. Saon, R. Sarikaya, G. Zweig: Improvements to the IBM hub 5e system, Proc. NIST RT-02 Workshop (2002)

30.20 G. Saon, G. Zweig, B. Kingsbury, L. Mangu, U. Chaudhari: An architecture for rapid decoding of large vocabulary conversational speech, Proc. Eurospeech, Vol. 3 (2003) pp. 1977–1981

30.21 S. Axelrod, V. Goel, B. Kingsbury, K. Visweswariah, R. Gopinath: Large vocabulary conversational speech recognition with a subspace constraint on inverse covariance matrices, Proc. Eurospeech, Vol. 3 (2003) pp. 1613–1616

30.22 S. Axelrod, R.A. Gopinath, P. Olsen: Modeling with a subspace constraint on inverse covariance matrices, Proc. Int. Conf. Spoken Lang. Process., Vol. 2 (2002) pp. 2177–2180

30.23 S. Wegmann, D. MacAllaster, J. Orloff, B. Peskin: Speaker normalization on conversational telephone speech, Proc. ICASSP, Vol. 1 (1996) pp. 339–342

30.24 M.J.F. Gales: *Maximum likelihood linear transformations for HMM-based speech recognition*, Vol. CUED/F-INFENG/TR291 (Cambridge Univ. Engineering Department, Cambridge 1997)

30.25 C.J. Leggetter, P.C. Woodland: Speaker adaptation of continuous density HMMs using multivariate linear regression, Proc. Int. Conf. Spoken Lang. Process., Vol. I (1994) pp. 451–454

30.26 S.F. Chen, J. Goodman: An empirical study of smoothing techniques for language modeling, Computer, Speech Lang. **13**(4), 359–393 (1999)

30.27 L.R. Bahl, P.V. deSouza, P.S. Gopalakrishnan, D. Nahamoo, M. Picheny: Robust methods for using context-dependent features and models in a continuous speech recognizer, Proc. ICASSP, Vol. I (1994) pp. 533–536

30.28 M. Padmanabhan, G. Ramaswamy, B. Ramabhadran, P.S. Gopalakrishnan, C. Dunn: Issues involved in voicemail data collection, Proc. DARPA Broadcast News Transcription and Understanding Workshop (1998)

30.29 L. Mangu, E. Brill, A. Stolcke: Finding consensus in speech recognition: Word error minimization and other applications of confusion networks, Computer, Speech Lang. **14**(4), 373–400 (2000)

30.30 E. Shriberg, A. Stolcke, D. Baron: Observations on overlap: Findings and implications for automatic processing of multi-party conversation, Proc. Eurospeech, Vol. 2 (2001) pp. 1359–1362

30.31 D. Povey, P. Woodland: Minimum phone error and I-smoothing for improved discriminative training, Proc. ICASSP, Vol. 1 (2002) pp. 105–108

30.32 D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, G. Zweig: FMPE: Discriminatively trained features for speech recognition, Proc. ICASSP, Vol. 1 (2005) pp. 961–964

30.33 M. Padmanabhan, M. Picheny: Large-vocabulary speech recognition algorithms, IEEE Comput. **35**(4), 42–50 (2002)

30.34 Google Desktop Developer Group: http://www.google.com/apis/ (2007)

30.35 B.E.D. Kingsbury, N. Morgan, S. Greenberg: Robust speech recognition using the modulation spectrogram, Speech Commun. **25**(1–3), 117–132 (1998)

30.36 M. Ostendorf, V.V. Digilakis, O.A. Kimball: From HMMs to segment models: A unified view of stochastic modeling for speech recognition, Proc. IEEE Trans. Speech Audio Process. **4**(5), 360–378 (1996)

30.37 J. Bridle, L. Deng, J. Picone, H. Richards, J. Ma, T. Kamm, M. Schuster, S. Pike, R. Regan: An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition, Final Workshop Report, Center for Language and Speech Processing (The Johns Hopkins University, Baltimore 1998)

30.38 G. Zweig, M. Padmanabhan: Dependency modeling with Bayesian networks in a voicemail transcription system, Proc. Eurospeech, Vol. 3 (1999) pp. 1335–1338

30.39 J. Bilmes: Buried Markov models, Proc. ICASSP, Vol. 2 (1999) pp. 713–716

30.40 M. Padmanabhan: Use of spectral peak information in speech recognition, Proc. NIST Speech Transcription Workshop (2000)

30.41 Ö. Çetin, M. Ostendorf: Multi-rate and variable-rate modeling of speech at phone and syllable time scales, Proc. Int. Conf. Acoust. Speech Signal Process., Vol. 1 (2005) pp. 665–668

30.42 M.P. Cooke, P.D. Green, L.B. Josifovski, A. Vizinho: Robust automatic speech recognition with missing and uncertain acoustic data, Speech Commun. **34**, 267–285 (2001)

30.43 S. Dharanipragada, M. Padmanabhan: A nonlinear unsupervised adaptation technique for speech recognition, Proc. Int. Conf. Spoken Lang. Process., Vol. IV (2000) pp. 556–559

30.44 R. Balchandran, R. Mammone: Non-parametric estimation and correction of non-linear distortion in speech systems, Proc. ICASSP, Vol. II (1998) pp. 749–752

30.45 H. Erdogan, R. Sarikaya, Y. Gao, M. Picheny: Semantic structured language models, Proc. Int. Conf. Speech Lang. Process., Vol. II (2002) pp. 933–936

30.46 R. Sarikaya, Y. Gao, M. Picheny: Word level confidence measurement using semantic features, Proc. ICASSP, Vol. I (2003) pp. 604–607

30.47 J. Bellegarda: Exploiting latent semantic information in statistical language modeling, Proc. IEEE **88**(8), 1279–1296 (2000)

30.48 F. Jelinek, C. Chelba: Putting language into language modeling, Proc. Eurospeech, Vol. 1 (1999) pp. KN-1–KN-4

30.49 I. Gurevych, R. Malaka, R. Porzel, H.P. Zorn: Semantic coherence scoring using an ontology, Proc. HLT-NAACL (2003) pp. 88–95

30.50 A. Likhododev, Y. Gao: Direct models for phoneme recognition, Proc. ICASSP, Vol. 1 (2002) pp. 89–92

30.51 V. Vapnik: The support vector method, Proc. Int. Conf. Artif. Neural Networds (1997) pp. 263–271

30.52 S. Della Pietra, V. Della Pietra, J. Lafferty: Inducing features of random fields, IEEE Trans. Pattern Anal. Mach. Intell. **19**(4), 380–393 (1997)

30.53 V. Venkataramani, W. Byrne: Lattice segmentation and support vector machines for large vocabulary continuous speech recognition, Proc. ICASSP, Vol. 1 (2005) pp. 817–820

30.54 L. Miller, M. Escabi, H. Read, C. Schreiner: Spatiotemporal receptive fields in the lemniscal auditory thalamus and cortex, J. Neurophysiol. **87**, 516–527 (2001)

30.55 J.G. Fiscus: A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), Proc. IEEE Workshop Autom. Speech Recognition Understanding, Santa Barbara (1997) pp. 347–355

30.56 Y. Freund, R.E. Schapire: Experiments with a new boosting algorithm, Proc. ICML (1996) pp. 148–156

30.57 O. Siohan, B. Ramabhadran, B. Kingsbury: Constructing ensembles of ASR systems using randomized decision trees, Proc. ICASSP, Vol. 1 (2005) pp. 197–200

30.58 IBM Research Communication Dept.: http://www.research.ibm.com/bluegene (2007)

30.59 IBM Research Communication Dept.: http://www.research.ibm.com/cell (2007)