

Toward Domain-Independent Conversational Speech Recognition

Brian Kingsbury, Lidia Mangu, George Saon, Geoffrey Zweig, Scott Axelrod,
Vaibhava Goel, Karthik Visweswariah, and Michael Picheny

IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

{bedk, mangu, gsaon, gzweig, axelrod, vgoel, kv1, picheny}@us.ibm.com

Abstract

We describe a multi-domain, conversational test set developed for IBM's Superhuman speech recognition project and our 2002 benchmark system for this task. Through the use of multi-pass decoding, unsupervised adaptation and combination of hypotheses from systems using diverse feature sets and acoustic models, we achieve a word error rate of 32.0% on data drawn from voicemail messages, two-person conversations and multiple-person meetings.

1. Introduction

The goal of IBM's Superhuman speech recognition project [1, 2] is to develop a domain-independent speech recognition system that matches or exceeds human performance across the full range of possible application domains, acoustic conditions and speaker characteristics. To foster work toward this admittedly aggressive objective, we have defined a test set comprising conversational American English material drawn from a number of application domains and set a goal of achieving annual 25% relative improvements in word error rate on this test set.

In this paper, we describe the components of our "Superhuman" test corpus, then describe our 2002 benchmark system for the corpus. To deal with the broad range of material present in the test set, we employed a recognition strategy based on multiple passes of recognition interleaved with unsupervised acoustic model adaptation and on combination of recognition hypotheses from systems using disparate feature sets and acoustic models. We 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 Superhuman test corpus at various stages of processing.

2. A multi-domain 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 on it. 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, comprising 113 minutes of audio. The data are 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 Meeting corpus [3, 4], comprising 95 minutes of audio. The data are 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). 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 30 minutes 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.

cc2 34 minutes 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.

vm Test data from the IBM Voicemail corpus, comprising 52 minutes of audio. This material was previously reported on as the E-VM1 test set [5], and is a superset of the test data in the Voicemail Corpus Part I and Part II distributed by the LDC. Unlike the other tests, the voicemail data are conversational monologues. The acoustic quality of the data are 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 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.

3. Signal processing

The systems in this work use either Mel-frequency cepstral coefficient (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 vari-

ance of the spectral estimates. The final recognition feature set for all systems in this work are generated by concatenating raw features from nine consecutive frames and projecting to a 60-dimensional feature space. The projection is a composition of a discriminant projection (either linear discriminant analysis or heteroscedastic discriminant analysis [6]) and a diagonalizing transform [7, 8].

Prior to the projection to the final, 60-d recognition feature space, the raw features are normalized. Three 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 max. normalization of $c0$; and (3) side-based mean and variance normalization of all features except $c0$ and max. normalization of $c0$. In max. 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 computed only from speech frames.

4. 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),

$$F(\theta) = \log P(X_s | s, \theta) - \frac{\lambda}{2} |\theta| \log(N_s) \quad (1)$$

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 hours of Switchboard data and 18 hours of Callhome English 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 [9]. This lets us exploit technology developed for fast decoding of conversational speech [10] for fast MMIE training as well. The second system models leaves with SPAM (subspace precision and means) models [11, 12]. 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, the SPAM models do not directly use BIC-based model selection.

4.1. Canonical acoustic models

We use two feature-space transformations, vocal tract length normalization (VTLN) [13] and maximum-likelihood feature-space regression (FMLLR) [14], 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 $\pm 20\%$ 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) [14], 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 models.

4.2. Acoustic model adaptation

At test time, we also use (MLLR) adaptation [15] 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 non-speech 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.

5. Language modeling and recognition lexicon design

The data used to train the language models consist of 3M Switchboard words, 16M Broadcast News words, 1M Voice-mail words and 600K call center words. For the initial rescaling of the word internal lattices we used a 4-way interpolated language model, each of the components being a back-off 3-gram LM using modified Kneser-Ney smoothing [16]. The mixture weights (0.45*Swb + 0.25*BN + 0.15*VM + 0.2*CC) are optimized on a held-out set containing 5% of each of the training corpora. For the final rescaling of the left-context lattices the 3-gram mixture components are replaced with 4-gram lan-

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

Table 1: Perplexities and OOV rates across different test sets

guage models, keeping the mixture weights the same. The 34K word vocabulary used in our experiments consists of all the high-count words from our training corpora. The pronunciation dictionary consists of 37K entries, yielding a ratio of 1.09 pronunciations per word in the vocabulary. Table 1 shows the perplexities and the out-of-vocabulary (OOV) rates for each of the five test sets.

6. Recognition process and performance

Recognition of data using the 2002 IBM Superhuman speech recognition system proceeds as follows:

- P1 Speaker-independent decoding. The system uses mean-normalized MFCC features and an acoustic model comprising 4078 left context-dependent states and 171K mixture components. Decoding is performed using IBM’s rank-based stack decoding technology [17].
- 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 comprising 4440 left context-dependent states and 163K 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 comprising 3688 word-internal context-dependent states and 151K 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 comprising 10437 left context-dependent states and 623K mixture components. This system uses max.-normalization of c_0 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 comprising 10450 left context-dependent states and 589K mixture components. This system uses mean normalization of all raw features.
- D A SPAM MFCC system comprising 10133 left context-dependent states and 217K 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 comprising 10441 left context-dependent states and 600K mixture components. This system uses max.-normalization of c_0 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 [18], 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 [19], 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 2.

7. 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 “Switchboard is representative of the acoustic-phonetic and stylistic properties” of conversational American English [20].

Multi-pass decoding with unsupervised adaptation and combination of disparate systems are effective techniques

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

Table 2: Word error rates (%) for the components of the 2002 Superhuman 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.

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 amount of computation incurred. Had we used only the MMIE PLP system and performed consensus decoding instead of confusion network combination in P7, the overall error rate on the test would increase to 33.1%.

8. Future work

Because the Superhuman project is expected to continue for the rest of the decade, we must be concerned about “training on the test set.” That is, through continued benchmarking and optimization on one test set, we may overspecialize our recognition systems. We believe that the broad range of material included in the test set mitigates, but does not eliminate, this problem. We therefore plan to add new components to the test set over the course of the project. This year we will begin benchmarking on an hour of material from the MALACH (multilingual access to large spoken archives) project [21]. The MALACH material is drawn from interviews with Holocaust survivors, and contains a high incidence of emotional speech, accented speech, age-related coarticulations, and disfluencies.

9. Acknowledgments

We are grateful to the International Computer Science Institute for providing us with the meeting data.

10. References

- [1] M. Padmanabhan and M. Picheny, “Towards superhuman speech recognition,” in *Proc. ASR Workshop*, 2000.
- [2] J. Huang, M. Picheny, and B. Ramabhadran, “Multi-domain robust speech recognition,” in *Proc. DARPA SPINE Workshop*, 2001.
- [3] A. Janin *et al.*, “The ICSI Meeting corpus,” in *ICASSP*, 2003.
- [4] <http://www.icsi.berkeley.edu/speech/mr>
- [5] M. Padmanabhan *et al.*, “Automatic speech recognition performance on a voicemail transcription task,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 433–442, October 2002.
- [6] G. Saon *et al.*, “Maximum likelihood discriminant feature spaces,” in *ICASSP*, 2000.
- [7] R. A. Gopinath, “Maximum likelihood modeling with Gaussian distributions for classification,” in *ICASSP*, 1998.
- [8] M. J. F. Gales, “Semi-tied full-covariance matrices for hidden Markov models,” Tech. Rep. CUED/F-INFENG/TR287, Cambridge University Engineering Department, 1997.
- [9] J. Huang *et al.*, “Improvements to the IBM Hub 5e system,” in *Proc. NIST RT-02 Workshop*, April 2002.
- [10] G. Saon *et al.*, “An architecture for rapid decoding of large vocabulary conversational speech,” submitted to *Eurospeech*, 2003.
- [11] S. Axelrod *et al.*, “Large vocabulary conversational speech recognition with a subspace constraint on inverse covariance matrices,” submitted to *Eurospeech*, 2003.
- [12] S. Axelrod, R. A. Gopinath, and P. Olsen, “Modeling with a subspace constraint on inverse covariance matrices,” in *ICSLP*, 2002.
- [13] S. Wegman *et al.*, “Speaker normalization on conversational telephone speech,” in *ICASSP*, 1996.
- [14] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” Tech. Rep. CUED/F-INFENG/TR291, Cambridge University Engineering Department, 1997.
- [15] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.
- [16] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” *Computer Speech and Language*, vol. 13, no. 4, pp. 359–393, 1999.
- [17] L. R. Bahl *et al.*, “Robust methods for using context-dependent features and models in a continuous speech recognizer,” in *ICASSP*, 1994.
- [18] M. Padmanabhan *et al.*, “Issues involved in voicemail data collection,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
- [19] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.
- [20] E. Shriberg, A. Stolcke, and D. Baron, “Observations on overlap: Findings and implications for automatic processing of multi-party conversation,” in *Eurospeech*, 2001.
- [21] B. Ramabhadran, J. Huang, and M. Picheny, “Towards automatic transcription of large spoken archives - English ASR for the MALACH project,” in *ICASSP*, 2003.