RECENT EXPERIMENTS IN LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION

J. Billa, T. Colhurst, A. El-Jaroudi, R. Iyer, K. Ma, S. Matsoukas, C. Quillen, F. Richardson, M. Siu, G. Zavaliagkos, H. Gish.

BBN Technologies
Cambridge MA 02138
jbilla@bbn.com

ABSTRACT

This paper describes the improvements that resulted in the 1998 Byblos Large Vocabulary Conversational Speech Recognition (LVCSR) System. Salient among these improvements are: improved signal processing, improved Hidden Markov Model (HMM) topology, use of quinphone context, introduction of diagonal speaker adapted training (DSAT), incorporation of variance adaptation in the MLLR framework, improvements in language modeling, increase in lexicon size and combination of multiple systems. These changes resulted in about a 7% absolute reduction in word error rates on a balanced Switchboard/Callhome English test set.

1. INTRODUCTION

The large vocabulary conversational speech recognition (LVCSR) task consists of recognition of natural spontaneous speech between speakers who may or may not be familiar with one another. This unconstrained conversational speech along with the resulting dysfluencies results in much higher error rates than more constrained tasks. Typically, on a balanced Switchboard/Callhome English test set, word error rates (WERs) are about 40%. The DARPA/NIST LVCSR evaluations present a common benchmark for the evaluation of research and technology within the LVCSR community. This paper reviews the 1998 Byblos LVCSR system which participated in the 1998 DARPA/NIST LVCSR evaluation.

First, we give a brief overview of the Byblos system and the various data sets used in training and testing of the system. Following this we review various experiments that resulted in improved performance over the 1997 system [13]. These improvements include our efforts to improve signal processing, acoustic and language modeling. Also we describe our experiences with the ROVER system combination algorithm [3].

Amro El-Jaroudi is Associate Professor of Electrical Engineering at the University of Pittsburgh, George Zavaliagkos is currently Distinguished Scientist at Lernout and Hauspie Speech Products.

2. SYSTEM DESCRIPTION

2.1. Signal Processing

The September 1998 LVCSR system uses a single, 45-dimensional feature stream. Features are extracted from overlapping frames of audio data, each 25ms long, at a rate of 100 frames per second. Each frame is windowed with a Hamming window, and an LPC smoothed, VTL warped log power spectrum is computed for the frequency band 125-3750 Hz. From this, 14 Mel-warped cepstral coefficients are computed. We use a gender-dependent, 128 term Gaussian mixture model, to compute a maximum-likelihood VTL warp parameter [12, 13]. To determine gender, we use a second Gaussian mixture to estimate a gender-independent VTL warp, and decide gender by thresholding this estimated stretch. The mean cepstrum and peak energy of each conversation is removed non-causally from the appropriate subvector. In addition, the feature vectors are scaled and translated so that, for each conversation side, the data has zero mean and unit variance. These base cepstral features with their first and second derivatives, together with the zeroth, first, and second derivatives of frame energy, compose the final 45-dimensional feature vector.

2.2. Acoustic Modeling

The acoustic feature stream is modeled in a gender-dependent manner with two pairs of HMM's, one pair for unadapted decoding and one pair for adapted decoding. Preliminary decoding passes use a phonetic tied-mixture (PTM) model, while the final decoding pass uses a state-clustered tied-mixture (SCTM) model. In both models, the atomic HMM is a 5-state chain with a minimum duration of 2 frames, and an output distribution that is a mixture of diagonal Gaussians (512 Gaussians per mixture in the PTM system, 80 per mixture in the SCTM system). Clustering is employed so that different HMM states may share the same distribution or the same codebook. The PTM system has 53 codebooks and 12,000 distributions, while the SCTM system has 3000 codebooks and 25,000 distributions. Both

the speaker-independent (SI) PTM and SCTM models are trained on 136 hours of Switchboard data and 15 hours of Callhome data.

The speaker-adapted models (SA), used in adapted decoding, are created by estimating for each training speaker a set of 256 diagonal transformation matrices; the components of each matrix are chosen so as to maximize an auxiliary function calculated during a prior forward-backward pass, as dictated by the EM algorithm. Once the transformation matrices are estimated for all speakers, the means and variances of the SA model are re-estimated to further improve the auxiliary function. This entire procedure is repeated three times to generate the final SA model [9, 1]. The same 151 hours of raw audio data is used for both the SI and the SA training.

2.3. Language Modeling and Recognition Lexicon

Two grammars are used at various phases of recognition. To create the lattice and N-best list, we use a trigram grammar on 35K words.

This grammar is trained from (i) all conversations of the Callhome English data (0.3 million words) (ii) all of the Switchboard data (3.1 million words), with the exception of the 1995, 1996 and 1997 evaluation sets, and (iii) 141M words of CNN with each article weighted by its similarity to the Switchboard and Callhome training. The second grammar is used for rescoring the N-best list (the scores from the grammar are interpolated to generate the final ordering of the list). The grammar uses a part of speech (POS) smoothing mechanism to interpolate the CNN data to the Switchboard and Callhome training data [6].

The lexicon comprises all non-name words seen in the Callhome data, together with all words seen in the Switchboard data, with the exception of the 1995, 1996 and 1997 evaluation sets, plus 10K additional words selected from the CNN data most similar to Switchboard.

2.4. Recognition

Decoding is done in five steps: (a) a speaker's gender and VTL parameter are estimated with Gaussian mixture models; (b) transcriptions are generated with the SI models; (c) MLLR adaptation mean and variance parameters are computed from these (errorful) transcriptions; (d) new N-best transcriptions are generated with adapted SA models; (e) more powerful language models are applied to rescore the N-best list and yield the 1-best transcription.

2.5. Data Sets

The various systems presented here were trained on one of two data sets. The first, Minitrain97, is a gender balanced 20 hour subset of the Switchboard corpus. The second, Evaltrain98 is composed of the entire Switchboard corpus and the Callhome English corpus. All systems were tested on a gender balanced subset of the 1997 Switchboard-II/Callhome evaluation test set.

3. EXPERIMENTS IN LVCSR

3.1. Vocal Tract Length Normalization (VTLN) and Signal Processing

VTLN is a transformation based on the premise that a singular reason for speaker feature variability is the differing vocal tract lengths of speakers. Differences in vocal tract length result in an apparent expansion or compression of the frequency axis as observed in formant trajectories. VTLN seeks to compensate for this variation in formant location by a warp of the frequency axis such that formant locations remain stationary across speakers. Last year, we had presented a maximum-likelihood VTLN (ML-VTLN) [12, 13] that was a significant improvement over our earlier formant based VTLN approach. The ML-VTLN approach uses a Gaussian mixture model (GMM) against which speakers were scored at a multiplicity of warps. The warp that scored the highest likelihood was then taken to be the VTLN stretch factor for that speaker. One deficiency of this approach is that the GMM shows an inherent likelihood bias for cepstra at different warps. To compensate for this effect the determinant of the VTL transformation is estimated empirically per speaker and applied.

Also, to compensate for the variation in the dynamic range of the cepstral variance across speakers, the cepstra were normalized to unit variance on a per speaker basis in addition to the a per speaker based non-causal cepstral means subtraction.

Table 1 summarizes the various VTLN and signal processing experiments. These improvements together resulted in a overall 1.4% absolute reduction in WER.

System Frontend	WER
ML-VTLN (initial baseline)	54.33
w/ variance normalization (baseline)	53.68
+ global bias removal	53.18
+ per speaker bias removal	52.93

Table 1: WERs for VTLN and signal processing experiments on Minitrain97.

3.2. Acoustic Modeling

We experimented with several ideas to improve acoustic modeling within our system: The HMM topology was changed to allow for a minimum of two frames compared

to the minimum of three before. This year we have adapted quinphones as our context in our SCTM models. We have found that in small training data sets, quinphones models do not provide any additional advantage over triphone models. This is most likely due to the manner in which contemporary HMM systems are trained with tied clusters. To elaborate: consider a data set where one has determined that "X" clusters are appropriate based on the available training data size. If "X" is small the clusters are most likely to be triphone based despite the allowance for longer quinphone context. In fact for our small data set (Minitrain97), less than 1% of the resulting clusters used quinphone context. However moving to quinphones on a sufficiently large corpus such as the entire switchboard corpus, significantly improves the performance of the system as is evident from results summarized in Table 2.

System	WER
Baseline (triphone)	48.35
+ min. 2-frame HMM topology	47.94
+ quinphones	46.61

Table 2: Experiments HMM topology and quinphones on Evaltrain98.

An addition to the Byblos system this year was the use of diagonal transformations in Speaker Adaptive Training (SAT). Previously Byblos made use of full-matrix adaptive training [9, 1]. This year we employed diagonal transformation matrices. We found that a 9 level regression class tree with diagonal matrices yielded performance almost equal to a 3 level deep tree with full matrixes (0.1% WER degradation). But diagonal matrices commute in such a way as to allow transformations to effectively take place in feature space. This permits a much less expensive implementation of SAT. Another possible approach would have been to employ 'constrained model space' SAT as proposed by Gales [4], which is more efficient than our traditional SAT for similar reasons.

MLLR Variance adaptation [4] also made its first appearance in Byblos this year. We employed diagonal transformations using a 3-level deep regression class tree which matched the regression classes used for MLLR means adaptation. It is possible to employ a deeper regression class tree for variances, but we were unable to profit by doing so. Diagonal variance adaptation is exceedingly easy to implement, and provided a 0.5% absolute reduction in WER.

3.3. Language modeling and Recognition Lexicon

Lexicons and language models can be improved by combining the sparse domain-dependent text with large amounts of out-of-domain data [8]. In the 1998 Byblos system, we incorporate the Broadcast News (BN) data with existing Switchboard and Callhome training text to: (i) improve the lexicon by reducing out-of-vocabulary (OOV) rates, and (ii) improve both the decoding and rescoring n-gram language models.

System	WER
SAT (3-level) +MLLR mean adapted	47.91
SAT (3-level) +MLLR mean and variance adapted	47.31
DSAT (9-level) +MLLR mean and variance adapted	47.44

Table 3: Experiments with DSAT and MLLR variance adaptation.

Typically, multiple recognition errors can be corrected for each OOV word recognized correctly, so it is useful to expand the recognition lexicon to reduce OOV rates. However, large lexicons also increase word confusability and recognition search costs, so it is important to choose the added words carefully. In this year's system we select words after pooling similarity weighted multi-domain data [7, 8] to both replace the infrequently observed words in the existing lexicon and expand the lexicon beyond the words observed in the domain-dependent training.

Similarity-weighted multi-domain text is used to estimate the decoding n-gram language models, providing improved training for existing and new words in the expanded lexicon. In addition, we use a more powerful language model, specifically a variation of a part-of-speech (POS) grammar that smooths multi-domain n-gram distributions [6, 8], to rescore N-best lists and yield the 1-best transcription. Interpolation weights used in the POS grammar are estimated on held-out in-domain Switchboard and Callhome training text.

Table 3.3 reports lexicon and language modeling gains obtained from using the multi-domain text. Three language models are referred to in Table 3.3: (i) M0 refers to a regular trigram language model trained only with the Switchboard and Callhome training text, (ii) M1 refers to a similarity-weighted trigram language model trained with added BN data, and (iii) M2 refers to the POS-smoothed grammar using word and POS n-gram distributions from all three domains.

3.4. System Combination via ROVER

Among the more interesting developments in LVCSR technology is Fiscus' ROVER algorithm [3] for system combination. ROVER is an algorithm whereby multiple systems can be combined to yield WERs lower than any of the constituent systems alone. We have improved on the original mechanism with the addition of a weighted selection procedure which is optimized to reduce WER. The key to large

Lexicon Size	Model	OOV Rate (%)	WER (%)
25 K	M0	2.7	47.9
35 K	M0	2.1	48.3
35 K	M1	2.1	46.6
35 K	+ M2	2.1	46.1

Table 4: Lexicon and language modeling gains for the Byblos98 system on an in-house development test set. "+" indicates model used for rescoring Nbest from the system described in the previous line.

reductions in WER with ROVER is the presence of systems with similar WERs but with dissimilar errors. In our experiments we have realized the greatest reduction in WER by combining the quinphone system output for test set analysis at three different frame rates (80,100 and 125 frames/sec) along with the corresponding triphone system. A weighted combination of these systems resulted in an overall 1.2% absolute reduction in WER.

System	WER
Quinphone, 80fps (qph-80)	48.75
Quinphone, 100fps (qph-100)	47.14
Quinphone, 125fps (qph-125)	49.20
Triphone, 100fps (tph-100)	47.86
ROVER with qph-80,100,125 and tph-100	46.10
Weighted ROVER with these four systems	45.90

Table 5: System combination with modified ROVER

4. REFERENCES

[1] T. Anastasakos, J. McDonough, and J. Makhoul. Speaker adaptive training: A maximum likelihood approach to speaker normalization. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, April 1997. IEEE.

[2] J.J. Godfrey et. al. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, San Francisco, March 1992. IEEE.

[3] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER). In *Proceeding of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–354, Santa Barbara, 1997.

[4] M. J. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. Technical Report CUED/F-INFENG/TR. 291, Cambridge University, Engineering Department, Cambridge, England, 1997.

[5] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[6] R. Iyer and M. Ostendorf. Transforming out-of-domain estimates to improve in-domain language models. In *Proceedings of EUROSPEECH-97*, volume 4, pages 1975–1978, Rhodes, September 1997.

[7] R. Iyer, M. Ostendorf, and H.Gish. Using out-of-domain data to improve in-domain language models.

[8] Rukmini Iyer. *Improving and Predicting Performance of Statistical Language Models in Sparse Domains*. PhD thesis, Electrical Engineering Department, Boston University, Boston, 1998.

[9] J. McDonough, T. Anastasakos, G. Zavaliagkos, and H. Gish. Speaker-adapted training on the switchboard corpus. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 1043–1046, Munich, April 1997. IEEE.

[10] L. Nguyen, R. Schwartz, F. Kubala, and P. Placeway. Search algorithms for software-only real-time recognition with very large vocabularies. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 91–95, Princeton, March 1993.

[11] M. Siu, H. Gish, and F. Richardson. Improved estimation,evaluation and application of confidence measures for speech recognition. In *Proceedings of EUROSPEECH-97*, Rhodes, September 1997.

[12] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalization on conversational telephone speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, May 1996.

[13] G. Zavaliagkos, J. McDonough, D. Miller, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Siu, and H. Gish. The BBN Byblos 1997 large vocabulary conversational speech recognition system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Seattle, May 1998. IEEE.