The 2000 BBN Byblos LVCSR System

Thomas Colthurst, Owen Kimball, Fred Richardson, Han Shu,
Chuck Wooters, Rukmini Iyer, Herbert Gish

GTE/BBN Technologies
70 Fawcett St., Cambridge Ma 02138
thomasc@bbn.com

ABSTRACT

This paper describes the 2000 BBN Byblos Large Vocabulary Continuous Speech Recognition (LVCSR) system. We briefly outline the training and decoding procedures used in the system, and explain in detail the new features we have added to the system in the past year. These new features include multiple adaptation stages, parallel path rescoring, and a new word confidence system. Word error rate results for all of these additions are presented for Hub-5 English test sets containing both Switchboard II and CallHome speakers.

1. Introduction

The 2000 BBN Byblos LVCSR system aims for state of the art performance on decoding spontaneous, conversational telephone speech. In particular, it is explicitly designed to achieve low word error rate (WER) on the NIST sponsored Hub-5 evaluations, the test sets of which consist of equal parts of Switchboard II and CallHome conversations [1]. In the March 2000 Hub-5 evaluation, Byblos achieved a WER of 29.1%.

This paper has two parts. First, we describe the current state of the Byblos system by describing the decoding procedure and the training methods used to create its various models. Then, we focus on the experiments we have run on the system over the past year, including those which led to improved performance over the previous Hub-5 Evaluation system [9]. These experiments include work on parallel path models, multiple adaptation stages, and a new training method for the word confidence system. For all of these experiments, error rates will be reported on DevSet98, a subset of the 1998 Hub-5 evaluation test set consisting of 7 Switchboard II and 7 CallHome conversations. (Each conversation contains two speakers and is approximately five minutes in length). The improvements described in this paper collectively lowered the WER on this development set by 2.1 %.

2. System Description

2.1. Signal Processing

Analysis in the Byblos system starts by breaking up the audio data into overlapping frames, each 25 msec long, at a rate of 100 frames per second (f/s). Each frame is windowed with a Hamming function, and an LPC smoothed, vocal track length (VTL) and Mel-warped log power spectrum is computed for the frequency band 125-3750 Hz. From this, the first 14 cepstral coefficients and a frame energy are retained; these are then normalized by non-causal mean cepstrum and peak energy subtraction and finally scaled and translated so that for each conversation side, the resulting feature vector has zero mean and unit variance in each dimension. For historical and performance reasons, analysis is run slightly differently in the gender and VTL warp estimation stages than it is in decoding stage [12].

2.2. Decoding

Decoding takes a collection of audio files (typically 8 kHz, 8-bit μ-law encoded) as input, along with a list of utterance start and end times and a speaker label for each utterance. These inputs are then put through the following stages: gender estimation, VTL warp estimation, analysis, unadapted decoding, adapted decoding, word confidence generation, and finally, system combination. The output is a transcription of each utterance, along with start and end times and a confidence measure for each of the utterance's words.

Gender Estimation The audio is analyzed at VTL warps 0.92, 0.94, …, 1.06. For each warp, each speaker's cepstra is scored with a Gaussian mixture model (GMM). Speakers with best scoring warps of 0.98 and above are labelled as female and the rest as male.

VTL Warp Estimation The audio is analyzed at VTL warps 0.88, 0.90, …, 1.12. For each warp, each speaker's cepstra is scored with a gender dependent (GD) GMM. Further, each speaker's 0.88 and 1.12 warped cepstra are scored against another GD GMM to estimate the Jacobian compensation factor for these warps; the compensation factors for the other warps are then linearly interpolated [9]. Finally, each speaker is given the VTL warp with the highest normalized score.

Analysis The cepstra to be used in decoding is produced next using for each speaker the VTL warp estimated in the previous stage. The cepstra and energy's first and second derivatives are added to the feature stream at this time, resulting in a 45 dimensional feature vector for each audio frame.

Unadapted Decoding In unadapted decoding, multiple search passes are performed across successively narrower search spaces of possible transcriptions of the audio data; later passes use more detailed acoustic and language models but use the results of earlier passes to constrain their search.

Specifically, the first pass uses a forward fast match search with non-crossword, phonetically tied mixture (PTM) acoustic models and a bigram language model. The second and third passes perform respectively backward and forward beam searches with approximate trigram language models and the same acoustic models; at the end of the third pass, a word lattice is created. This lattice is then searched in the fourth pass with crossword, state-clustered tied mixture (SCTM) acoustic models and a trigram language model; it produces an N-best list of the top 100 ranked possible transcriptions. Finally, this N-best list is rescored with parallel path acoustic models and a cross domain, part of speech (POS) smoothed language model.

Adapted Decoding Three stages of adapted decoding are performed, each identical to the unadapted decoding stage described above except in the acoustic models used and in the lack of parallel path rescoring. Each stage uses the rescored 1-best transcriptions of a speaker's utterances produced by the previous stage as the basis for maximum likelihood, linear regression (MLLR) adaptation. The MLLR transforms are estimated with 3 iterations of EM. The first and third adaptation stages both start with diagonal transformation Speaker Adaptive Trained (DSAT) acoustic models [9], but the first adapts them using 4 transformations and the third using 8 transformations. The second adaptation stage starts with the same speaker independent (SI) acoustic models used in unadapted decoding, but adapts them with 8 transformations.

Word Confidence Generation For every word in the 1-best rescored output of the third adapted decode, twenty-one features are gathered as inputs for a generalized linear model (GLM) [5]. Features include (in rough order of importance): the word's frequency in the N-best list, the utterance's language model score, the likelihood of the trigram ending at the word, the number of phonemes in the word, the utterance's normalized (divided by number of frames) acoustic score, the utterance's average signal power, the number of words in the sentence, and binary, "am-I-that-word" features for the words AND, IF, IS, IT, THAT, THEY, WHERE, and WOULD. The output of the GLM is the word's confidence score.

System Combination To generate alternative hypotheses for system combination, the very last two stages of the third adapted decode, the lattice transcription and LM rescoring stages, along with the word confidence generation step, are repeated with cepstra analyzed at 125 and 80 frames per second. (Everything else, including the acoustic models, remains the same). The outputs from these stages, along with the output of the normal, 100 frame per second, third adapted decode, comprise the systems to be combined.

The combination is done following a modified ROVER algorithm [3]. First, the outputs of the three systems are aligned, and then the possible words for each alignment slot are voted upon. A word's vote is equal to its confidence score from a system times that system's predetermined weight, summed over all the systems it occurs in, plus a bonus if the word is hypothesized by a majority of the systems. The winning words form the final output of the decode; confidences for the winning words are generated by a GLM with the word's votes as features.

2.3. Training

Acoustic Training Acoustic training takes 120 hrs of Switchboard and 17 hours of CallHome data, labels it using forced phonetic alignment with simple bootstrapped models, uses the labels to grow separate decision tree clusters for both the Gaussians and their mixture weights (the five state HMM transition probabilities are unclustered), initializes the Gaussians via the k-means algorithm, and trains all the parameters with five passes of the EM algorithm. This process is done for both the quinphone SCTM and triphone PTM models, both of which are gender dependent. In the end, the PTM models contains 53 Gaussian clusters and 12,000 mixture weight clusters, with 512 Gaussians per mixture, while the SCTM models have 3,000 Gaussian clusters and 25,000 mixture weight clusters, with 80 Gaussians per mixture.

To create the DSAT models used in the first and third adapted decoding passes, a set of 256 diagonal transformation matrices are estimated for each training speaker, and then both the means and variances of the Gaussians are re-estimated. Both estimations are done with the EM algorithm so as to maximize the likelihood of the joint transformation; the entire procedure is repeated three times to create the final DSAT models.

The parallel path HMM [4] attempts to model the distinct segment-length trajectories of a phonetic unit using separate HMM paths. In this work, we used only two path models, each with a conventional 5-state, left-to-right topology. The model is initialized from training data in which each phonetic segment is given a path label based on a bootstrap segment model (here, a segmental k-means labeler). The path labels are combined with state labels from our conventional (non-parallel-path) HMM and used to initialize 1-Gaussian-per-state models used in clustering. State clustering is accomplished using a modification of the usual tree-based, divisive clustering in which parallel states (states in the same position on different paths of a model) start out together in the root of the tree. In the process of generating the next binary split in the clustering tree, in addition to the usual splitting along phonetic contexts with linguistic questions, we allow parallel states to be split or not based on a path question. In this way, although the complete paths are always kept distinct (we don't allow them to merge or cross), if two parallel states of the path have very similar statistics, their observation distributions can be shared. The total number of Gaussians used in the parallel-path model was kept the same as the number used in the system's SCTM models. Following clustering, the model is trained using the normal HMM forward backward algorithm.

The GMMs used in both gender and VTL estimation are estimated as follows: first, all unwarped cepstra are used to create a single, 256 Gaussian mixture model. Next, for each warp, each training speaker's warped cepstra are scored against this GMM. A new GMM is created from each speaker's best scoring cepstra, and the process is repeated three times.

Language Model Training Four different language models are used by the Byblos system. The first two, a bigram and approximate trigram, are used by the forward and backwards decoding passes respectively, and are trained from three million words of Switchboard and eighty thousand words of CallHome. The trigram used in the lattice pass is trained from this data plus 140 million words of CNN which is weighted to reflect the domain mismatch. Finally, the trigram grammar used in N-best rescoring is trained from all of this data, but the three domains' contributions are part of speech (POS) smoothed, with interpolations weights estimated on held out test data.

The 35,000 word dictionary was created from all the non-name words in the CallHome and Switchboard data, plus 10,000 additional words from the CNN data which were selected as the most similar to those in Switchboard.

Confidence & System Combination Training The word confidence systems and system combination weights are trained by performing a complete decode on a development (held-out) set of data for which the truth is known. (The weights used to reorder N-best lists after rescoring are also trained via the development decode.) Each system's GLM is maximum likelihood (ML) trained, with a Bayesian Information Criterion (BIC) used to guide a greedy search over possible feature sets. There are 140 available features, including binary, "am-I-that-word" features for the most frequent 100 words in the acoustic training.

The trained GLMs are then used to generate word confidences for the development decodes, and the system combination weights are numerically optimized using Powell's method to maximize WER [10].

3. Experiments & New Features

3.1. Multiple Adaptation Stages

This is the first year in which Byblos runs multiple adaptation stages; previous attempts to use more stages with SAT models showed no gain. In [11], Dragon suggested that the alternating use of different models can give better results. We find this to be true, and alternating between SAT and SI models gives us a 1.6% net gain (see Figure 2).

Jack-knifing the adaptation transcripts is also advocated in [11]; that is breaking up a speaker's utterances into multiple chunks, and then decoding a given chunk with models adapted to all of the other chunks' transcriptions. We find that jack-knifing with seven chunks definitely hurts later stages and helps early ones only slightly if at all (see Figure 3).

Confidence System	NCE
Old GAM/NMI system	0.161
New GLM/BIC system	0.174
GLM + squares	0.175
GLM + word features	0.180
GLM + word & phoneme features	0.180

Figure 3: Confidence System Improvements

3.2. Parallel Path Rescoring

Parallel-path rescoring is also new to the Byblos evaluation system this year. Previous work with a triphone-based parallel-path system showed a .9% WER improvement for combining the parallel-path with a triphone conventional HMM in N-best rescoring on DevSet98. The quinphone parallel-path model estimated for this evaluation gives us a .3 WER improvement on the unadapted stage. Although we do not know the reason for this smaller gain, this was the first time the parallel-path was run with quinphones and we expect larger improvements with further investigation.

3.3. New Confidence System

In previous years, word confidences were generated by a generalized additive model (GAM) which was trained to maximize normalized mutual information (NMI); feature selection was done by a greedy algorithm evaluated on a held out test set [5] [8].

The new confidence system uses GLM models (a proper subset of the class of GAM models) which are ML trained using a BIC criteria for feature selection. In addition, several new features, including binary "am-I-that-word" features for the most frequent 100 words, were added. Figure 3 summarizes the improvements in normalized cross entropy (NCE), the measure NIST uses to compare confidences between systems, for these changes. It also reports on two changes which did not improve NCE and thus did not make it into the final system: adding squares of all the (non-binary) features (so as to allow the GLM to more closely approximate a GAM), and "how-many-times-do-I-contain-that-phoneme" features. It should be noted that in both of these changes, the feature selector did choose some of the new features (such as N-best frequency squared or occurrence counts of DH or ER); the resulting GLMs merely did not increase the NCE.

3.4. Silence Modeling

Our normal SCTM clustering algorithm allocates many mixture models to quinphones with silence as their center phone. Figure 4 describes results of using only one mixture model for such quinphones. In this experiment, only forty hours of training data were used for the acoustic models in these results. The .4 gain indicates that the large number of silence Gaussian mixtures in the baseline system represents a suboptimal allocation of parameters. Since this experiment was run after the evaluation, this feature was not present in the evaluation system.

Pass	Model	Adapt?	# Xfmations	BW WER	LT WER	Opt WER	Rescore WER
1	SI	no	n/a	53.6	43.4	42.6	42.1
2	SAT	yes	4	45.4	38.6	37.9	37.4
3	SI	yes	8	42.2	36.9	36.4	36.4
4	SAT	yes	8	41.3	36.6	36.1	36.0
5	SI	yes	8	41.5	36.1	35.6	35.8

Figure 1: Gains on Multiple Adaptation Passes

Pass	Model	Jack-Knifed?	# Xfmations	BW WER	LT WER	Opt WER	Rescore WE R
2	SAT	no	4	45.4	38.6	37.9	37.4
2	SAT	yes	4	48.0	38.9	38.2	37.1
4	SAT	no	8	41.3	36.6	36.1	36.0
4	SAT	yes	8	47.5	38.0	37.2	36.9

Figure 2: Jack-Knifing Results

System	WER
40 hr. train	47.5
”” + only 1 silence model	47.1

Figure 4: Results of Using Only One Silence Mixture Model

3.5. Per-Utterance LM Adaptation

The idea here is that given separate Switchboard and CallHome language models, we can adapt a LM interpolation parameter to an utterance by maximizing

$$\sum_{i=1}^{100} \lambda P(s_i|\text{Switchboard}) + (1-\lambda)P(s_i|\text{CallHome})$$

with s_i the sentences in that utterance's N-best list, and then using $\lambda_{\text{max}}P(\cdot|\text{Switchboard}) + (1-\lambda_{\text{max}})P(\cdot|\text{CallHome})$ to rescore the s_i.

This idea was tried in two different ways. First, standard trigram language models analogous to those used in the lattice pass were trained for Switchboard and CallHome. Performing per-utterance LM adaptation with these models gave a .26% gain over the lattice pass WER, but combining the per-utterance LM adapted scores with the cross domain, POS smoothed LM scores reduced this gain to only a .02% gain over using only the latter.

In the second implementation, cross domain, POS smoothed language models were trained for Switchboard and CallHome and used for per-utterance LM adaptation. (Note that both of these language models were trained from Switchboard, CallHome, and CNN data, but for the Switchboard model, they were combined so as to minimize the perplexity of the LM on a held out Switchboard II test set). This gave only a 0.05% gain over the normal rescoring language model WER; see figure 5. Although we believe this is an interesting and promising method, due to the lack of improvement in the complete system, it was not included in the evaluation system.

Pass	WER
Lattice	42.51
Lattice + per-utterance	42.24
Rescoring	42.08
Rescoring + per-utterance	42.05
Per-utterance w/ POS LM	42.03

Figure 5: Per-Utterance LM Adaptation Results

4. References

1. J. J. Godfrey et. al, “SWITCHBOARD: Telephone speech Corpus for research and development”, *Proc. ICASSP-92*, San Francisco, March 1992.
2. J. McDonough, T. Anastaskaos, G. Zavaliagkos, H. Gish, “Speaker-Adapted Training on the Switchboard Corpus”, *Proc. ICASSP-97* Munich, Germany, April 1997.
3. J. G. Fiscus. “A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER),” in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347-354, Santa Barbara, 1997.
4. R. Iyer, O. Kimball, and H. Gish, “Modeling Trajectories in the HMM Framework,” *Proc. EUROSPEECH-99*, Sept. 1999.
5. T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.
6. Rukmini Iyer. *Improving and Predicting Performance of Statistical Language Models in Sparse Domains.* PhD thesis, Electrical Engineering Department, Boston University, Boston 1998.
7. L. Nguyen, R. Schwartz, F. Kubala, and P. Placeway. “Search algorithms for software-only real-time recognition with very large vocabularies.” In *Proc. of the*

ARPA Human Language Technology Workshop, pages 91-95, Princeton, March 1993.

8. M. Siu, H. Gish, and F. Richardson. “Improved estimation, evaluation, and application of confidence measures for speech recognition,” *Proc. EUROSPEECH-97*, Rhodes, September 1997.
9. J. Billa, *et al*, “Recent Experiments in Large Vocabulary Conversational Speech Recognition”, *Proc. ICASSP-99*, Phoenix, May 1999.
10. F. Acton. *Numerical Methods that Work*, pages 464-467, Harper and Row, New York, 1970.
11. B. Peskin *et al*, “Improvements in Recognition of Conversational Telephone Speech”, *Proc. ICASSP-99*, Phoenix, May 1999.
12. P. Dognin, A. El-Jaroudi, J. Billa, “Parameter Optimization for Vocal Tract Length Normalization,” *Proc. ICASSP-2000*, Istanbul, May 2000.