THE BBN BYBLOS 1997 LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION SYSTEM

G. Zavaliagkos, J. McDonough, D. Miller, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Siu, H. Gish

GTE/BBN Technologies, 70 Fawcett Street, Cambridge, MA 02138
gzaval@bbn.com

ABSTRACT

This paper presents the 1997 BBN Byblos Large Vocabulary Speech Recognition (LVCSR) system. We give an outline of the algorithms and procedures used to train the system, describe the recognizer configuration and present the major technological innovations that lead to performance improvements. The major testbed we present our results for is the Switchboard Corpus, where current word error rates vary from 27% to 34% depending on the test set. In addition, we present results on the CallHome Spanish and Arabic tests, where we demonstrate that technology developed on English Corpora is very much portable to other problems and languages.

1. INTRODUCTION

This paper presents the 1997 BBN Byblos Large Vocabulary Conversational Speech Recognition (LVCSR) system, with emphasis on our work on the Switchboard Corpus. Switchboard consists of spontaneous conversations of speakers unknown to one another on a prescribed topic ([1]). Word error rates (W.E.R) for Switchboard has dropped considerably in the last few years, from the high 70's in 1992 to around or below 30% in 1997.

We first give a brief overview of the Byblos system (Section 2), including a description of the signal processing techniques, the training algorithm and the language modeling procedures. The we focus more on Switchboard and the latest innovations that lead to improved performance (Section 3); in particular we describe our recent experience with signal processing and vocal tract normalization issues, speaker adaptation, our efforts to combine out of-domain text to enhance language modeling and, finally, we present some results on the effects of the size of the training set to performance.

Also in this paper we present comparative results across languages for the CallHome Corpora (English, Spanish and Arabic). The CallHome corpora consist of conversations between family members and/or friends conversing freely in their native language over the telephone; one side of the phone call is in the U.S. and the other abroad (which also introduces issues with noisy lines). Furthermore, because the two callers know each other, speech is *much* more spontaneous than Switchboard, and we have the additional problem of the frequent use of foreign (to the language in question) words.

The summary of our experiments with the English CallHome Corpus (Section 4) is that it is quite harder to recognize than Switchboard. When comparing English CallHome with foreign CallHome we find that technology improvements are largely language independent.

2. SYSTEM DESCRIPTION

2.1. Signal Processing

The 1997 Byblos 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 ([2]) warped log power spectrum is computed for the frequency band 125-3750 Hz. From this, 14 Mel-warped cepstral coefficients are computed. These coefficients together with the frame energy and their first and second derivatives compose the raw 45-dimensional feature vector.

The feature vectors are normalized in several ways before being used for training or decoding. The mean cepstrum and peak energy of each conversation is removed noncausally from the appropriate sub-vector. In addition, the feature vectors are scaled and translated so that, for each gender, the pooled training data has zero mean and unit variance.

To do the processing described above, we require knowledge of the gender of each speaker and an estimate of a vocal tract length (VTL) parameter for that speaker. We use a gender dependent, 128 term Gaussian mixture model, to compute a maximum likelihood VTL warp parameter. To determine gender, we use a second Gaussian mixture to es-

A. El-Jaroudi is currently with the university of Pittsburgh; J. McDonough is currently at Johns Hopkins University

timate a gender independent VTL warp, and decide gender by thresholding this estimated stretch.

2.2. Acoustic models

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. The atomic HMM is a 5-state chain with a minimum duration of 3 frames, and an output distribution that is a mixture of diagonal Gaussians (256 Gaussians per mixture (phoneme) in the PTM system, 40 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 8.5-13.5K Gaussians (depending on the number of phonemes in the underlying language) and 12K distributions, while the size of the SCTM system is determined by the amount of training data available. For English (with 140 hrs of speech available) it has 3K codebooks (120K Gaussians) and 25K distributions, while for Arabic (20hrs of training speech) it has 1K codebooks (40K Gaussians) and 15K distributions.

Estimation of these HMM's occurs in multiple steps, typically involving the k-means algorithm to generate the Gaussians and the EM algorithm to generate the mixture weights and re-estimate the Gaussians. Between EM iterations, the codebooks are adjusted to prevent under-trained Gaussians: those with fewer than 10 samples are merged with neighboring ones, and MAP smoothing is used to interpolate the variances of individual Gaussians with the pooled codebook variances.

The speaker-adapted model (SA) ([3]), used in adapted decoding, is created by estimating for each training speaker a set of four 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. The same data is used for both the SI and the SA training.

2.3. Decoding

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 matrices 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.

Steps (b) and (d) are done in a nearly identical fashion. A first pass over the test data does a fast match to produce scores for numerous word endings using the PTM model ([4]). A second (forward) and third (backward) pass using the PTM model and an approximate trigram grammar generate a lattice including trigrams and crossword expansions. Next, a fourth pass with the SCTM model produces 1-best and word-dependent N-best transcriptions.

2.4. Language modeling

Three different grammars are used at various phases of recognition. To create the lattice and N-best list, we use a trigram grammar created from all available training conversations (3.5M words for English, considerable less for Spanish and Arabic). The lexicon comprises all words seen in the aforementioned training data (25658 for English).

For Switchboard, two other grammars are used for rescoring the N-best lists. The first is a variable 5-gram grammar ([7]) and the second grammar is made by adding articles from the CNN text database to the Switchboard trigram in a weighted fashion according to part-of-speech (POS) similarity ([6]).

3. EXPERIMENTS ON SWITCHBOARD

3.1. VTL Normalization

The VTL transformation is motivated by the fact that formant frequencies for different speakers lie in different places due to differences in the vocal tract length. Therefore, stretching the frequency axis to account for vocal tract length can make the resulting spectra look more similar across speakers. The Maximum Likelihood (ML) VTL estimation starts from the following observation: if the best VTL stretch was known, we could apply the named warp to each speaker's data and then build a Gaussian Mixture model out of the resulting pool of warped data. But of course we do not know the best warp to start with. To emulate the same process we start by bootstrapping the mixture model with the unwarped speaker's data, and then use ML to tell us what the most likely warp is.

We began by choosing an arbitrary pool of 60 conversations per gender. Each speaker was analyzed with 13 different warps in the range of 0.88 to 1.12, and used the Gaussian Mixture to select the most likely one. The newly warped data was used to re-estimate the model, and the process was repeated 2 more times. The final model was then used as the "VTL warp generator" for all other training as well as test speakers. The results obtained with this technique are presented in Table 1 for 18 hours of training speech, and compared with a formant based VTL estimator as presented

in ([5]). As we can see, ML-VTL clearly outperforms formant based VTL.

VTL method	Train	Test	% W.E.R
-	-	-	45.7
formant	y	y	44.3
ML	-	y	42.5
ML	y	y	41.7

Table 1: Results with VTL

Given the good results obtained by ML-VTL we explored with little success a number of potential improvements over the basic method. The main issues tackled here is whether there is any advantage in estimating VTL based on only voiced frames, and on whether one could improve performance by having more than one VTL parameter (one for voiced and one for unvoiced frames). We also looked at the effect of the quantization of the VTL warp, by refining the initial selected stretch by decreasing the search step from 0.2 to 0.02 around the best estimate. Results appear on table 2.

VTL method	% W.E.R
baseline ML-VTL	41.7
refined quantization	41.6
voiced frames only	41.8
two warps for V/UV	41.6

Table 2: Refined VTL estimation

3.2. Signal Processing

We experimented with a number of ideas in signal processing, namely non-causal versus causal CMS and energy normalization, hamming versus Blackman windows, analysis bandwidth, frame size and different CMS for speech and non-speech frames. The baseline includes causal CMS and energy normalization, Blackman window, .3-3.3KHz bandwidth, 20ms frame size. The results are summarized in Table 3. The system that combined all signal processing improvements gave a 2.7% absolute gain over the baseline, which is smaller than the sum of the parts, but still a very significant gain.

3.3. Language Modeling

The goal of this work is to enhance the baseline Switchboard trigram. Two other grammars are estimated and then used for rescoring the N-best list (the scores from these grammars are interpolated to generate the final ordering of the list). The first is a variable 5-gram grammar obtained by collecting counts for all 5-grams in the data, and then selectively pruning back to lower-order grammars based on

method	% W.E.R
baseline	44.26
1. Non causal CMS	43.78
2. Hamming window	43.74
3. 125-3750Hz BW	42.07
4. 25ms frame	43.72
5. Speech/nspeech CMS	44.52
1-4 combined	41.56

Table 3: Signal Processing Improvements

the similarity between the n- and (n-1)-grams. The second rescoring grammar is made by adding articles from the CNN text database (141M words) to the Switchboard trigram in a weighted fashion. A separate weight is assigned to each CNN article based on the similarity between its part-of-speech (POS) n-grams and Switchboard. The final trigram is then smoothed with the likelihood of POS class given word history. Results appear on Table 4

Baseline	41.1
+ POS Smoothing with CNN	39.9
+Variable 5-gram	39.4

Table 4: Rescoring with enhanced language models

We also explored the effects of Lexicon building on performance: with a 10K lexicon the OOV rate was 2.32% and the error rate 40.2%; using a 25K lexicon reduced the OOV rate to 1.38% and the error rate to 39.3%. As we can see, we got an almost 1 to 1 reduction in error rate and OOV rate. This roughly agrees with empirical observations by many researchers in WSJ, that one corrects 1.2 errors for every OOV word that is covered. The ratio is smaller here because the error rate for Switchboard is much higher.

We also see a small gain (0.5%) from segmenting the training conversation transcriptions at major punctuation marks (. ; ? and !) instead at speaker turns, as was done in the past.

3.4. Effects of training size

The purpose of this work was to quantify the effect of the training data size for both acoustic and language modeling training (Table 5).

The empirical fit for the language modeling training is that there is a 3.5% gain for every 10-fold increase in the training text size. For acoustic modeling, an 8-fold increase in the training size gave a 5% improvement in performance; however, at the current operating point, the performance seems to saturate quicker with increases in acoustic rather than language training size.

training size	% W.E.R
18 hrs of speech	44.2
60 hrs	40.4
140 hrs	39.3
LM training size	% W.E.R
70K words	47.0
170K	45.6
2M	42.3
3M	41.7

Table 5: Training size and Performance

3.5. Summary of improvements

In the previous four sections we presented a number of individual improvements, with some of them evaluated on small training data conditions. When we trained on all 140 available hours of speech we found that the gains were mostly additive. The total gain for the language model improvements was 3.4% and the gain for combining the VTL normalization with other signal processing improvements was 5.5%. Overall, our performance improved from 36.2% in 1996 to 27.3% in 1997. Performance on the evaluation set provided by NIST was slightly worse, at 35.5%. The main factor that accounted for this degradation is that the test set was from the Switchboard-II collection, which happened 3 years later than Switchboard. As a result the language model coverage was very poor: the list of topics the speakers were involved with was very different, the OOV rate jumped from 0.7% to 1.7%, and 0.4% of the test words was so hard that human transcribers decided to mark them as unintelligible (they were still counted as errors though).

CallHome versus Switchboard: In the 1997 LVCSR evaluation we were also tested on a CallHome English test set, and the error rate there was 53.7%. As we mentioned in the introduction, CallHome tests are particularly harder because of the very spontaneous nature of speech and the presence of noise on the international phone side; furthermore, because the speakers are familiar to one another and use specific jargon (sometimes even in a foreign language), the OOV rate is much higher, at around 3.8%. To indicate how much harder the test was, 1.8% of the words were marked as unintelligible by human transcribers; we estimate that just these unintelligible words contributed 4.5% to the total error.

4. MULTI-LINGUAL EXPERIMENTS

We have found that the major technological improvements that reduced word error rates for the Switchboard Corpus (continuous densities, VTL, speaker adaptation and speaker adaptive training) improve performance across all languages we have tested so far. In other words, state-of-the art speech recognition technology is mostly language independent. For example, Table 6 gives the relative gains for speaker adaptation on the development sets across the three languages:

	English Switchboard	Spanish	Arabic
No adapt	39.2%	64.1%	63.3%
SI adapt	36.1%	61.1%	-
SAT adapt	34.5%	59.3%	58.6%

Table 6: Adaptation improvements across languages

Table 7 gives statistics on the training data size and word error rates for the three CallHome languages, English, Spanish, and Arabic, using the NIST Spring 1997 and Fall 1997 evaluation test sets. Although we see that performance varies some from language to language, we believe that most of the difference is accounted for by the different amounts of data available and statistical variability. For example, if we used the empirical rules presented in section 3.4, and assumed that equal amount of data was available as for English (140 hrs of speech and 3.5M words), we would expect that we would gain 4.9% for the additional speech and 3.9% for the additional language text (crossreffrence with table 5), which would bring the performance from 61.6% down to 52.8%, which is almost exactly where English currently is.

Language	Available speech	LM text	W.E.R
English	140hrs	3.5M words	53.7%
Spanish	56hrs	876K	57.9%
Arabic	17hrs	174K	61.6%

Table 7: Performance across languages for CallHome tests.

5. REFERENCES

[1] J.J. Godfrey et. al., "SWITCHBOARD: Telephone speech Corpus for research and development" *Proc. ICASSP-92*, San Francisco, March 1992.

[2] S. Wegmann et. al., "Speaker Normalization on Conversational Telephone Speech", *Proc. ICASSP-96* Atlanta

[3] J. McDonough, T. Anastasakos, G. Zavaliagkos, H. Gish, "Speaker-Adapted Training on the Switchboard Corpus", *Proc. ICASSP-97* Munich, Germany, April 1997.

[4] L. Nguyen, R. Schwartz, F. Kubala, P. Placeway, "Search Algorithms for Software-Only Real-Time Recognitions with Very Large Vocabularies" *???*

[5] E. Eide et.al., "A Parametric Approach to Vocal Tract Length Normalization", *Proc. ICASSP-96* Atlanta, May 1996

[6] R. Iyer and M. Ostendorf, "Transforming Out-of-Domain Data to Improve In-Domain Language Models", *Proc. EUROSPEECH-97*, Rhodes, Greece, 1997.

[7] M-H. Siu and M. Ostendorf, "Variable N-gram Language Modeling and Extensions for Conversational Speech", . *Proc. EUROSPEECH-97*, Rhodes, Greece, 1997.