RECENT IMPROVEMENTS IN THE CU SONIC ASR SYSTEM FOR NOISY SPEECH: THE SPINE TASK

Bryan Pellom and Kadri Hacioglu

Center for Spoken Language Research
University of Colorado at Boulder
{pellom,hacioglu}@cslr.colorado.edu

ABSTRACT

In this paper we report on recent improvements in the University of Colorado system for the DARPA/NRL Speech in Noisy Environments (SPINE) task. In particular, we describe our efforts on improving acoustic and language modeling for the task and investigate methods for unsupervised speaker and environment adaptation from limited data. We show that the MAPLR adaptation method outperforms single and multiple regression class MLLR on the SPINE task. Our current SPINE system uses the Sonic speech recognition engine that was recently developed at the University of Colorado. This system is shown to have a word error rate of 31.5% on the SPINE-2 evaluation data. These improvements amount to a 16% reduction in relative word error rate compared to our previous SPINE-2 system fielded in the Nov. 2001 DARPA/NRL evaluation.

1. INTRODUCTION

The Speech in Noisy Environments (SPINE) task attempts to measure and inspire improvements in state-of-the-art processing for robust continuous speech recognition [1]. The task has several challenges: limited task-dependent training data (~20 hours), multiple military noise environments in both training and testing, unsegmented audio streams as well as a limited amount of speech per task session for recognizer adaptation.

In November of 2000 (SPINE-1) and November 2001 (SPINE-2) the Naval Research Laboratory (NRL) evaluated systems on the task with support by DARPA. Participating sites in the 2001 evaluation included: SRI, IBM, University of Washington, University of Colorado (CU), AT&T, the Oregon Graduate Institute (OGI), Mississippi State, ATR, and Carnegie Mellon University (CMU). Many of these sites have previously reported results on SPINE-1 [2-4] and SPINE-2 tasks [5-8]. The best performing systems on that task used adaptation in either the feature or model-domain and also included the use of multiple parallel speech recognizers trained from several feature types (e.g., MFCC, PLP, root cepstrum). Output from each recognizer is generally combined through a hypothesis fusion method to produce a single output that is lower than the error rates of any single recognizer (e.g., see [5,6]).

The University of Colorado participated in both SPINE-1 [4] and SPINE-2 evaluations. Our November 2001 system was for the first time based on the University of Colorado speech recognizer named Sonic [9]. During that evaluation our single best recognizer output had an official error rate of 37.5% at a decoding speed of 9 times real-time. In this paper, we describe recent improvements both in terms of general recognizer development and task-dependent modeling. We focus on issues related to lowering the error rate of our single-best recognizer fielded on the SPINE task and do not consider the issue of recognizer fusion in this work.

2. THE SPINE TASK

The SPINE task uses the ARCON Communicability Exercise (ACE) that was originally developed to test communication systems [10] and consists of collaboration between a pair of talkers who participate in a battleship simulation. One participant plays the role of a Firing Officer (e.g., controlling weapon systems such as a laser cannon and mines) while the other participant plays the role of a Search Officer (e.g., manning the radar and sonar equipment). Each player is situated in a separated sound isolated room and use military handsets and headsets that are appropriate for the simulated acoustic conditions. During the exercise, the two participants collaborate to search and destroy targets by declaring and confirming grid locations (x-axis & y-axis coordinates) to fire upon. The grid locations in SPINE-1 consisted of confusable words from the Diagnostic Rhyme Test (DRT). For SPINE-2, the grid points consisted of less confusable military words. For each booth, noise indicative of typical military environments is played through loud speakers. The SPINE-1 evaluation data considered six noise environments: aircraft carrier control decision center, AWACS airplane, a military vehicle, a military field shelter, an office environment, and a quiet environment. SPINE-2 extends on SPINE-1 data by considering the six noise types in addition to military tank and helicopter environments. The resulting noisy speech from each booth is recorded through head-worn microphones before being passed through a simulated communications channel. In this paper we consider only speech recognition on the non-coded speech channel.

3. THE SONIC ASR ENGINE

3.1. Current ASR System Architecture

Our most recent fielded evaluation system in November 2001 (SPINE-2) was designed using Sonic: The University of Colorado large vocabulary continuous speech recognition system [9]. Sonic is based on continuous density hidden Markov (CDHMM) acoustic models. Context dependent triphone acoustic models are clustered using decision trees. Each model has three emitting states with gamma probability density functions for duration modeling. Features are extracted as 12 MFCCs, energy, and the first and second differences of these parameters, resulting in a feature vector of dimension 39. The search network is a reentrant static tree-lexicon. The recognizer implements a two-pass search strategy. The first pass consists of

0-7803-7663-3/03/$17.00 ©2003 IEEE

a time-synchronous, beam-pruned Viterbi token-passing search. Crossword acoustic models and 3-gram or 4-gram language models (in an approximate and efficient way) are applied in the first pass of search. The first pass creates a lattice of word ends. During the second pass, the resulting word-lattice is converted into a word-graph. Advanced language models (e.g. dialog-act and concept based, long span) can be used to rescore the word graph using an A* algorithm or to compute word-posterior probabilities to provide word-level confidence scores.

Sonic provides an integrated environment that incorporates voice activity detection (VAD), speech enhancement as well as various feature and model-based adaptation and normalization methods. The recognition architecture provides support for rapid portability to new languages. In 2002, Sonic was ported from English to the Spanish, Turkish, and Japanese languages.

3.1. General Recent Improvements

Our SPINE-2 system in Nov. 2001 represented our initial implementation of the Sonic speech recognizer. The fielded system used a flat structured lexicon, class-based trigram language model consisting of manually determined word compounds, single regression iterative MLLR mean and global variance scaling transform, and generalized triphone acoustic models. Since Nov 2001, we included an efficient lexical tree search, integrated a decision tree triphone acoustic model trainer, added support for 4-grams into our first-pass search, implemented data-driven word compounding, and incorporated additional feature normalization (cepstral variance normalization, VTLN) and speaker adaptation (MAPLR adaptation) methods.

4. SPINE SYSTEM OVERVIEW

Our SPINE system consists of a novel integrated speech detection and multiple pass recognition search as shown in Figure 1. During each recognition pass, a voice activity detector (VAD) is dynamically constructed from the current adapted system acoustic models. The VAD generates a segmentation of the noisy audio into utterance units and LVCSR is performed on each detected speech region. The resulting output (a confidence tagged lattice or word string) is then used to adapt the acoustic model means and variances in an unsupervised fashion. The adapted acoustic models are then reapplied to obtain an improved segmentation, recognition hypothesis, and new set of adapted system parameters. The integrated adaptation procedure can be repeated several times resulting in sequential improvements to both segmentation and recognition hypotheses.

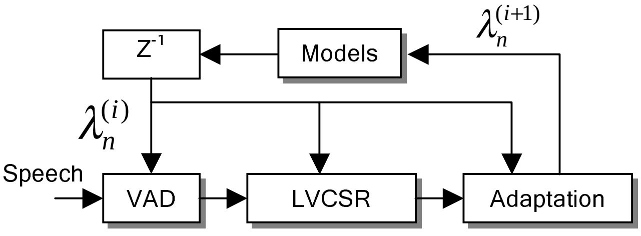

Figure 1: Diagram of SPINE multi-pass recognition search.

For the SPINE task, we have found that tight coupling between the segmentation and recognition system is essential for robust performance. Furthermore we illustrate how this integrated approach leads to simpler methods for voice activity detection for noisy environments. The following sections describe our current system for the SPINE task in detail.

4.1. Training Data

Acoustic and language model training data for the SPINE-2 evaluation consisted of conversations that were used for both training and testing in the previous SPINE-1 evaluation and conversation sides listed as training and development test for the SPINE-2 evaluation. For the SPINE-2 evaluation we optimized our recognizer settings on the provided 1.1-hour development test data before incorporating both the data and recognizer settings into our final system. Table 1 summarizes the training data used in the experiments described in this paper.

Training Data Source	Number of Utterances	Total Hours (Talk-Time)
SPINE-1 train	11,973	8.7
SPINE-1 eval.	12,079	7.3
SPINE-2 train	6,129	3.4
SPINE-2 dev.	1,941	1.1
Total	32,122	20.5

Table 1: SPINE-2 evaluation system training data

4.2. Acoustic Model

The acoustic trainer for Sonic is based on sequential estimation using Viterbi forced alignment and phonetic decision tree state clustering [12]. Alignments were initially boot-strapped using Wall Street Journal acoustic models. During Viterbi forced alignment we used a single MLLR mean and variance transform on the gender-dependent models to improve the alignment quality for each speaker session. After alignment, the models are estimated using decision tree state clustering and the procedure is repeated to obtain improved alignments and model parameter estimates. Our first-pass acoustic models consist of gender-dependent (within-word and cross-word) triphones using standard 39-dimensional MFCC features. Our second-pass (adaptation pass) acoustic models are normalized by both cepstral variance and vocal tract length [13].

.

4.3. Language Model

For the SPINE-2 evaluation in Nov. 2001 we developed a class N-gram language model trained from the 32k utterances shown in Table 1. This work was motivated by the fact that the grid-point labels were changed from SPINE-1 to SPINE-2 and class language models provided a convenient means for capturing the task specific word usage for targeting objects in the battleship game. Our class language model was based on 3 word classes: row (x-axis), column (y-axis), and name (user name). Words were grouped into row and column classes through inspection of the training data for SPINE-2. In this task there are several words that can be modeled as belonging to multiple classes. For example, the spoken words in the spelling of "VON" (read as "Victor Oscar Nancy") overlap with elements of the row grid axis class ('Victor' and 'Oscar' are part of the row class in SPINE-2). To deal with these ambiguities, we utilized a semi-automated tagging system originally developed for training class N-gram models for the DARPA Communicator task.

The task language model also contains word compounds for improved recognition. Our language model fielded in Nov. 2001 contained 115 compounds determined by manual inspection of the training data. In this paper we considered improving the existing SPINE language model by using the data driven method proposed in [14] for determining word compounds. This method uses the geometrical average of the direct and reverse-bigrams to determine candidate word compounds. Our current language model using data driven word compound clustering has a vocabulary of 1664 words and includes 180 word compounds.

4.4. Audio Segmentation

Our audio segmentation method iteratively estimates segment boundaries between adaptation passes and uses the adapted system acoustic models in decision-making. The segmenter consists of a 2-state (speech/non-speech) hidden Markov model that is dynamically constructed on each ASR adaptation pass. A speech state is constructed by combining the top 4 mixture components (by mixture weight) from the context-independent speech states of our decision tree clustered models. A silence state is constructed from all mixture components of non-speech context-independent states (e.g., breath, laughter, garbage, silence). The resulting HMM states (600 mixture components for speech, 288 mixture components for silence) are normalized such that the mixture weights sum to one. A Viterbi search is performed over each session using the 2-state HMM model. The speech / silence boundaries are determined through back-tracing the best path through the network. The segmentations are improved using 2 heuristics: (i) speech segments separated by less than 0.25 are merged, (ii) speech segments that are less than 0.10 seconds in duration are deleted. Finally, all speech segments are dilated by 0.25 seconds to avoid cutoff of weak fricatives and other low-energy sounds. We point out that this audio segmentation approach avoids the necessity of training separate speech/non-speech models and also avoids acoustic mismatch between VAD and system acoustic models in subsequent adaptation passes.

4.5. Acoustic Adaptation

In the SPINE task both speaker and environment variability are quite large. So, the adaptation of the speech recognizer to better match the test condition is crucial. To cope with such variability we have implemented several techniques that can be considered in two broad classes: feature-based and model-based techniques. In feature-based methods the observations, i.e. the feature vectors input to the speech recognizer, and in model-based methods the parameters of the acoustic models, i.e. HMM means and variances, are modified. Examples of feature-based normalization are cepstral mean subtraction (CMS), vocal tract length normalization (VTLN) and cepstral variance normalization. In CMS the long-term average of cepstral feature vectors is estimated and subtracted from the computed cepstral feature vectors. In VTLN, the best warping factor is determined by line searching over a range of values to maximize the likelihood of the adaptation data, given the recognized transcription. These processes are followed by feature variance normalization. These methods have been applied during both training and decoding in our SPINE system.

Model-based adaptation methods can be further categorized into two broad classes: direct and indirect. In direct adaptation, the HMM model parameters are directly adapted. However, in the indirect method a set of shared transformations are first estimated and then applied to the respective HMM models. Usually the maximum a posteriori (MAP) estimation is used for the direct method by incorporating some a priori knowledge to overcome data sparseness. In the indirect method the transformations are usually estimated in the maximum likelihood (ML) sense. A recent work in [15,16] unifies both methods in the MAP sense and demonstrates improved performance.

Several modes of adaptation are possible; supervised vs. unsupervised and block vs. incremental. In the unsupervised case, the transcription is not known and should be estimated in some form; either as a single best string or a word lattice. In incremental adaptation the models are adapted as enough data becomes available, and the new models are used to decode the incoming data, which, in turn, is used to readapt the models. In block adaptation, the adaptation is started after all data is available. We consider several adaptation schemes:

- Maximum likelihood linear regression (MLLR):
  (i) incremental / block, (ii) single class / multiple class,
  (iii) best string / word lattice
- Maximum a posterior linear regression (MAPLR):
  (i) block (ii) best string / word lattice
  (iii) regression class tree.

Our initial SPINE-2 system used a single class, block MLLR mean and variance transform using the best string from the speech recognizer tagged with confidence scores (word posterior probabilities) derived from a word graph. Despite some improvement in the Hub-5 task, extending from a single regression class to 6 classes degrades performance in the SPINE task. We believe this is due to the smaller amount of adaptation data in SPINE compared with Hub-5. This motivated us to work with a dynamic version of multiple class MAP adaptation using regression class trees. In the next section, we report performance gains obtained with more sophisticated adaptation techniques.

5. EVALUATION

The November 2001 SPINE-2 evaluation data consisted of 64 talker-pair conversations totaling 3.5 hours of stereo audio (2.8 hours of talk-time). On average, each of the 128 conversation sides contains 1.3 minutes (78 seconds) of speech activity.

5.1. Segmentation

Audio segmentation was evaluated by measuring the frame classification and word error rates for our baseline SPINE system when automatic and hand-labeled speech segments were used. Our baseline system uses single regression class MLLR mean and diagonal covariance transform. Results are shown in Table 1. We see that the voice activity detection method has an initial frame classification rate of 7.44% (Table 1a). After the first adaptation pass the segmenter produces fewer errors (final frame classification error rate of 6.93%) and the recognizer is better able to reject silence regions that have been misclassified as speech (e.g., the number of inserted words drops from 172 to 108). The word error rate difference between automatic and hand-segmented data is negligible (0.5% absolute).

Processing Stage	Automatic			Hand
	(a)	(b)	(c)	(d)
First-Pass	7.44%	172	41.8%	41.0%
MLLR-1	6.95%	108	33.9%	33.4%
MLLR-2	6.93%	112	33.2%	32.7%

Table 2: Segmentation performance summary. Results are shown for (a) speech/silence frame classification error rate; (b) number of inserted words during silence regions; (c) word error rate for automatic segmentation; (d) word error rate for hand-labeled segmentation.

5.2. Word Error Analysis

Table 3 summarizes word error rates (WER) across iterative adaptation passes and total real-time processing factors for several SPINE-2 system configurations. Iteration "0" in Table 3 refers to first-pass recognition. Real-time factors are measured on a single processor 1.7 GHz Intel Pentium 4 and include processing time incurred through automatic segmentation. Our baseline system without adaptation was found to have a 41.8% WER at 1.8x real-time. Furthermore incremental online adaptation based only on MLLR mean transformation provides nearly a 10% relative reduction in error with a modest cost in terms of processing speed. In fact, based on the general improvements listed in Section 3.1, this 1-pass incremental adaptation system compares favorably with our multiple pass Nov. 2001 system which has a 37.5% WER at 9x real-time.

The use of single regression class MLLR mean and diagonal covariance transforms iterated over 2 adaptation passes provides a considerable reduction in error (error drops from 41.8% to 33.2%). However, as many sites reported in both the 2000 and 2001 workshops, increasing to more than one transform generally degrades system performance perhaps due to lack of sufficient adaptation data (33.8% WER compared with 33.2% WER in Table 3d). Finally, the MAPLR algorithm using the single-best word-posterior probability weighted output provides a measurable reduction in error compared to the baseline of a single regression MLLR mean and variance transform. Further, the generalization of the technique to operate on the word-lattice representation provides an additional gain of 0.4% absolute. However, we point out that this improvement comes at a higher computational cost (e.g., 16.4x real-time compared with 6.4x).

System Description		Word Error Rate (%)			Real
		Iter 0	Iter 1	Iter 2	Time
(a)	Baseline, No Adapt.	41.8	--	--	1.8
(b)	Single Pass Inc. Adapt	37.7	--	--	2.0
(c)	1 Reg. Class MLLR	41.8	33.9	33.2	5.2
(d)	6 Reg. Class MLLR	41.8	34.2	33.8	4.6
(e)	Single-Best MAPLR	41.8	33.3	31.9	6.4
(f)	Lattice MAPLR	41.8	32.7	31.5	16.4

Table 3: Word error rate and real-time factor for SPINE-2 evaluation systems: (a): Baseline system without speaker adaptation; (b): system incorporating online incremental adaptation in a single pass; (c): single regression class MLLR with global variance scaling; (d): system using 6 MLLR regression classes; (e): word-posterior weighted single-best hypothesis MAPLR adaptation; (f): Lattice-based MAPLR.

6. CONCLUSIONS

The paper has presented several recent improvements to the University of Colorado (CU) SPINE-2 evaluation system. Our current implementation uses the newly developed CU Sonic ASR system. Our current best single recognizer system has an overall error rate of 31.5% at a real-time factor of 16.4. Comparatively, the single best recognizer based on MFCC features in [5] had a word error rate of 32.5% on the same evaluation set. We point out that the two best systems fielded in the 2001 evaluation had a real-time factor of 88 and 121 respectively. Based on these comparisons, we feel that the system presented in this paper represents the state-of-the-art in single recognizer performance on the SPINE-2 task.

7. REFERENCES

[1] T. Crystal, A. Schmidt-Nelson, E. Marsh, "Speech in Noisy Environments (SPINE) Adds News Dimension to Speech Recognition R & D," *Proc. HLT Conf.*, San Diego, March 2002.

[2] R. Singh, M. Seltzer, B. Raj, R. Stern, "Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination," *Proc. ICASSP*, Salt Lake City, 2001.

[3] R. Rose, H. Kim, D. Hindle, "Robust Speech Recognition Techniques Applied to a Speech in Noise Task," *Proc. Eurospeech*, Aalborg Denmark, 2001.

[4] J. Hansen, R. Sarikaya, U. Yapanel, B. Pellom, "Robust Speech Recognition in Noise: An Evaluation using the SPINE Corpus," *Proc. Eurospeech*, Aalborg Denmark, 2001.

[5] V. Gadde, A. Stolcke, D. Vergyri, J. Zheng, K. Sonmez, A. Venkataraman, "Building an ASR System for Noisy Environments: SRI's 2001 SPINE Evaluation System," *Proc. ICSLP*, pp. 1577—1580, Denver, Sept. 2002.

[6] C. Zheng, Y. Yan, "Run Time Information Fusion in Speech Recognition," *Proc. ICSLP*, pp. 1077—1080, Denver, Sept. 2002.

[7] O. Cetin, H. Nock, K. Kirchhoff, J. Bilmes, M. Ostendorf, "The 2001 GMTK-Based SPINE ASR System," *Proc. ICSLP*, Denver, Sept. 2002.

[8] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, R. Sarikaya, "Robust Speech Recognition in Noisy Environments: The 2001 IBM SPINE Evaluation System," *Proc. ICASSP*, pp. 53—56, Orlando, 2002.

[9] B. Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer", Technical Report TR-CSLR-2001-01, CSLR, University of Colorado, March 2001.

[10] http://www.arcon.com/ddvpc

[11] C. J. Legetter, and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech & Language, Vol. 9, pp. 171—185, 1995.

[12] W. Reichl, W. Chou, "Robust Decision Tree State Tying for Continuous Speech Recognition", IEEE Trans. on Speech and Audio Processing, Vol.8, No.5, Sept. 2000.

[13] L. Welling, H. Ney, and S. Kantahak, "Speaker Adaptive Modeling by Vocal Tract Normalization," IEEE Trans. on Speech and Audio Processing, Vol. 10, No. 6, pp. 415—426, Sept. 2002.

[14] G. Saon, M. Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition," IEEE Trans. on SAP, Vol. 9, No. 4, pp. 327—332, May, 2002.

[15] O. Siohan, C. Chesta, and C.-H. Lee, "Joint Maximum a Posteriori Adaptation of Transformation and HMM Parameters," IEEE Trans. on Speech & Audio Proc. Vol. 9, No. 4, pp. 417-428, 2001.

[16] O. Siohan, T. Myrvoll, and C.-H. Lee, " Structural Maximum a Posteriori Linear Regression for Fast HMM Adaptation", Computer, Speech and Language, 16, pp. 5-24, January 2002.