summary of deliverables for

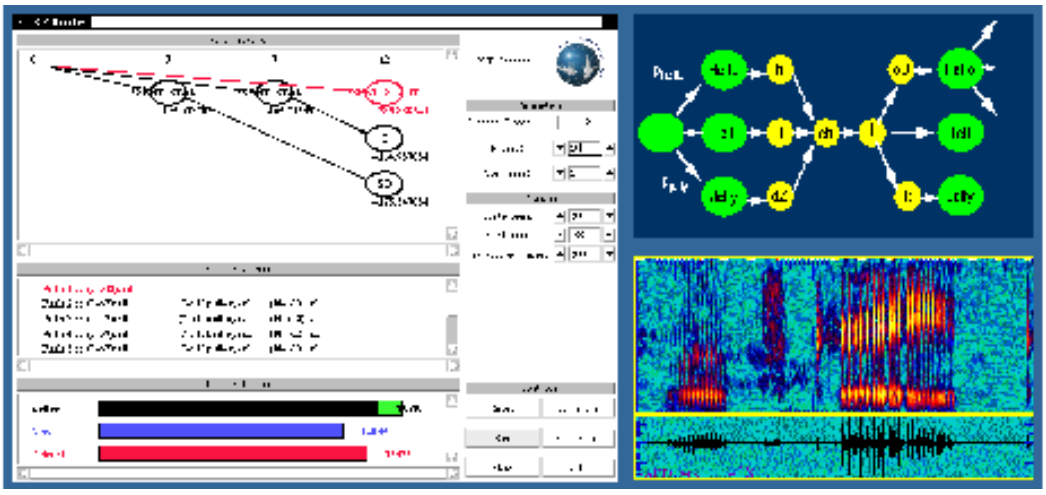
Robust Low Perplexity Voice Interfaces

Subcontract No. 43556

submitted to:

Dr. Fred J. Goodman
Signal Processing Center
The MITRE Corporation
1820 Dolley Madison Blvd., M/S W622
McLean, Virginia, USA 22102-3481

December 31, 2001



submitted by:

F. Zheng and J. Picone
Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University
Box 9571, 413 Simrall, Hardy Road
Mississippi State, Mississippi 39762
Tel: 662-325-3149, Fax: 662-325-3149
primary email contacts: {zheng, picone}@isip.msstate.edu

EXECUTIVE SUMMARY

This report summarizes the third phrase for our one-year collaboration with the MITRE Corporation on a project titled “Robust Low Perplexity Voice Interfaces.” The major effort for the third portion of this project was to again collaborate with MITRE on the development of a robust recognition technology for SPINE2 evaluation.

In the third phase of this project, we have successfully completed three milestones:

- Developed an HMM based triphone/trigram system for SPINE2 evaluation. This system achieved a WER of 56.9% on uncoded speech and a WER of 67.0% on coded speech in this evaluation.
- Delivered a Multiple-CPU Eval Package which allows a user to easily run complete experiments using multiple CPUs.
- Integrated an N-best utility to the prototype system to support N-best list output for our recognition server in the Communicator framework.

We participated in the SPINE2 evaluations for the second time this quarter. For this evaluation, we developed a single triphone/trigram system for both uncoded and coded speech. Results were 56.9% WER on uncoded speech, and 67.0% WER on coded speech. Our biggest problem was segmentation of the data, which cost us approximately 10% absolute in WER. Post-evaluation experiments have been conducted using reference segments to calibrate the effect of incorrect segmentation. We achieved a WER of 43.7% on uncoded speech and a WER of 55.1% on coded speech. From uncoded speech to coded speech, the performance degraded only ~10% absolute in WER.

In addition, the performance on “quiet” data was worse than the performance on “noisy” data, which concerned us. It seems that training models on the “quiet” condition from the clean data has the potential to improve the performance on “quiet” speech. Results on coded speech indicated that the MELP vocoder is very promising in narrowband communications, since MELP has the lowest bit rate at 2.4 kbps but results in the lowest relative degradation (9.3%) among provided four vocoders (>13.8%).

In this quarter, we also demonstrated the first use of our advanced technology, Support Vector Machines (SVM), on the SPINE application. A hybrid SVM/HMM system was developed that combines the advantages of HMMs and SVMs. In a preliminary experiment, the hybrid SVM/HMM system was used to rescore 10-best lists and delivered a 32.7% WER on a SPINE2 dry-run subset. Those 10-best lists were generated from a bigram word-internal HMM system with a WER of 33.2%. This slight improvement is not significant, but the hybrid SVM/HMM system has potential to improve speech recognition performance considering this system was highly tuned on Alphadigits task and only 10-best lists were not rich enough for potential improvement.

We have delivered a multiple-CPU evaluation package which allows a user to easily run complete experiments with training and decoding using multiple CPUs. A user can specify which computers are to be used, and configure the parameters from the command line. In addition, we have integrated an N-best list utility to the prototype system to support the N-best list output for the recognition server within Communicator framework. The output format of the recognition server was also standardized according to the request of Lockheed Martin.

TABLE OF CONTENTS

1. INTRODUCTION

In the past decade, state of the art speech recognition systems have achieved tremendous advances. However, military applications often place a higher demand on systems than cooperative uses, because the ambient environments in which military systems operate require robustness to extreme amounts of acoustic noise. Further, military communications need to transmit data over narrow bandwidth channels, and employ advanced speech compression algorithms to do so. Speech compression coding has a negative effect on the accuracy of a speech recognition system [1]. In order to achieve our goal of developing a deployable system, our collaboration with the MITRE Corporation focuses on the investigation of practical technologies in robust speech recognition for building fieldable military applications.

This quarter we have been mainly focusing on *The 2001 Speech in Noisy Environments Evaluation 2* (SPINE2) [2]. The SPINE2 evaluation is the second attempt to assess the state of the art and practice in speech recognition technology in noisy military environments. The goal of this evaluation is to promote research progress in this area, to provide the opportunity for participants to try out new ideas for developing robust speech recognition systems that are of both scientific and practical interest, and to measure the performance of this technology. Participation in SPINE2 in collaboration with MITRE was mainly a chance to demonstrate our practical techniques in robust speech recognition, to investigate the robustness of the system in noise environments, to challenge the impact of coded speech to the system, and ultimately to provide the fieldable tactical speech recognition technology.

Previously in this project, we have developed a real-time recognition system for the DARPA Resource Management (RM) database [3]. This task has a 1000 word vocabulary and a bigram perplexity of 60. Our baseline system achieves a WER of 3.4% running at 9.7 xRT. Our real-time system achieves a WER of 5.0% at 1 xRT on a 600 MHz Pentium processor. We have integrated this RM system into a recognition server that supports the DARPA Communicator framework [4].

2. SPINE2 EVALUATIONS

SPINE2 was similar to the SPINE1 evaluation held last year [5], but contained additional noise and vocoded speech conditions. We again collaborated with MITRE to participate in this evaluation. Last year, we submitted a system based on bigram lattice generation and word-internal triphone rescoring [6] which achieved a 56.2% WER. For SPINE2, we submitted a triphone/trigram system which achieved a 56.9% WER on uncoded speech and 67.0% WER on coded speech [7]. Considering this year's data was much harder than last year's, these numbers were respectable. A detailed description of this system is provided below.

2.1. SPINE2 Data Preparation

The official SPINE2 corpus is divided into training, dry-run, and evaluation sets, each containing uncoded and coded speech data. The source material for the training data consists of uncoded speech data from the SPINE1 training and evaluation sets and a new SPINE2 training set. These comprised a total of 324 conversation sessions which included about 33.6 hours of speech data from 64 speakers. The dry-run data consists of 32 uncoded speech sessions from 4 speakers. The

evaluation data consists of 64 uncoded speech sessions from 32 speakers. The audio corpus is derived from data consisting of 8 military noise conditions (Quiet, Office, Street, Car, Carrier, Bradley, F16 and Helo). The coded speech data sets are generated from the above conversations processed by one of four standard DoD vocoders [8]: LPC, CVSD, CELP and MELP.

The audio data in the SPINE corpus consists of two channels, each sampled at 16 kHz. This corpus was released in SPHERE files that employ a 16-bit linear sampled data format. These recordings consist of conversations between two speakers. Both sides of the conversation were recorded simultaneously. Speakers were immersed in different simulated noise environments, and also vocoded using different speech coders. The noisy environment was simulated by seating each speaker in a sound chamber in which a previously recorded military background noise environment was accurately reproduced. Though the speakers talk freely, the total vocabulary used is limited to about 6000 words.

For this evaluation, we downsampled all the data from 16 kHz to 8 kHz using the Matlab downsampling utility. The training and dry-run data were segmented using transcription segments. As did last year, segmentations for the evaluation data were not provided. We used our energy-based endpoint detector to broad segmentations. Since the data is extremely noisy, segmentation performance was poor. However, we did not have the time or resources to focus on this aspect of the problem.

The segmented raw data then underwent the same feature extraction procedure we used last year [6]. This involved generating 12 FFT-derived cepstral coefficients and log-energy. These features were computed using a 10 ms frame duration with a 25 ms Hamming window. First and second derivative coefficients of the base features were appended to produce a 39-dimensional feature vector. The 12 base cepstral feature were then debiased using side-based cepstral mean subtraction.

2.2. SPINE2 Training

For SPINE2 evaluation, we trained 16-mixture Gaussian HMM word-internal models and cross-word triphone models for two passes of decoding respectively. These triphone models used a 3-state left-to-right self-loop topology. The special topologies were used for the interword short

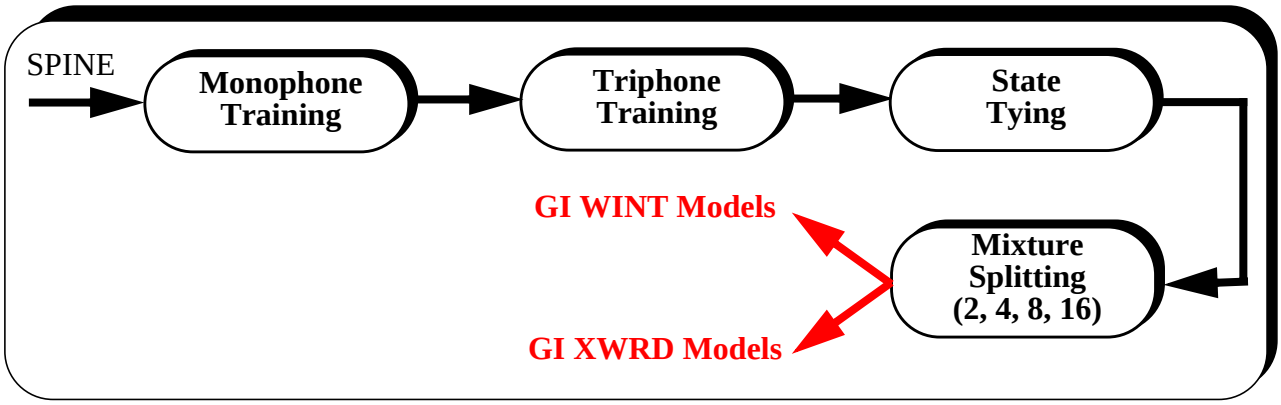


Figure 1. An overview of the training procedure used in the SPINE2 evaluation

pause (sp) model and the silence (sil) model respectively. The training procedure (shown in figure 1) involved first generating single-mixture monophone models. These monophone models were trained using the Baum-Welch algorithm. The trained monophone models were then used to seed word-internal or cross-word context dependent (CD) triphone models. A state-tying procedure was used to cluster those states and models that were statistically similar. The state-tied triphone models were iteratively trained from one mixture up to 16-mixture Gaussians per state.

For this task, we assumed we didn't know the information whether the data was uncoded or coded, so a single system was trained for both uncoded speech and coded speech. All the official released uncoded and coded training data were used to train the models for dry-run experiments. Eventually dry-run data was added into training data to train the final models for the evaluation.

2.3. SPINE2 Decoding

Recognition is performed using the ISIP time-synchronous Viterbi decoder [9]. The decoding procedure for SPINE2, shown in Figure 2, was performed in two stages. Stage 1 used the 16-mixture word-internal triphone models and a bigram language model to generate word lattices. VTLN was performed before Stage 2. The best VTLN warping factor for each speaker was determined on a per conversation side basis. Using the longest utterance from each speaker, we performed a forced alignment for each warping factor with the hypothesis from stage 1. The warping factor which gave the maximum likelihood score for speech frames was used for generating the rest of the VTLN feature data for that speaker. In stage 2, we rescored the lattices generated in stage 1 using 16-mixture cross-word triphone models and a trigram language model with the VTLN-warped features. The output of stage 2 is the final hypothesis string of the system.

In the SPINE evaluations, participants are encouraged to use a standard lexicon and trigram language model provided by CMU [10]. This lexicon has a vocabulary size of 5720 words. The trigram LM contained 54,073 trigrams, 22,748 bigrams, and 5,720 unigrams. A bigram LM was generated from the CMU trigram LM by discarding the trigram portion of the LM.

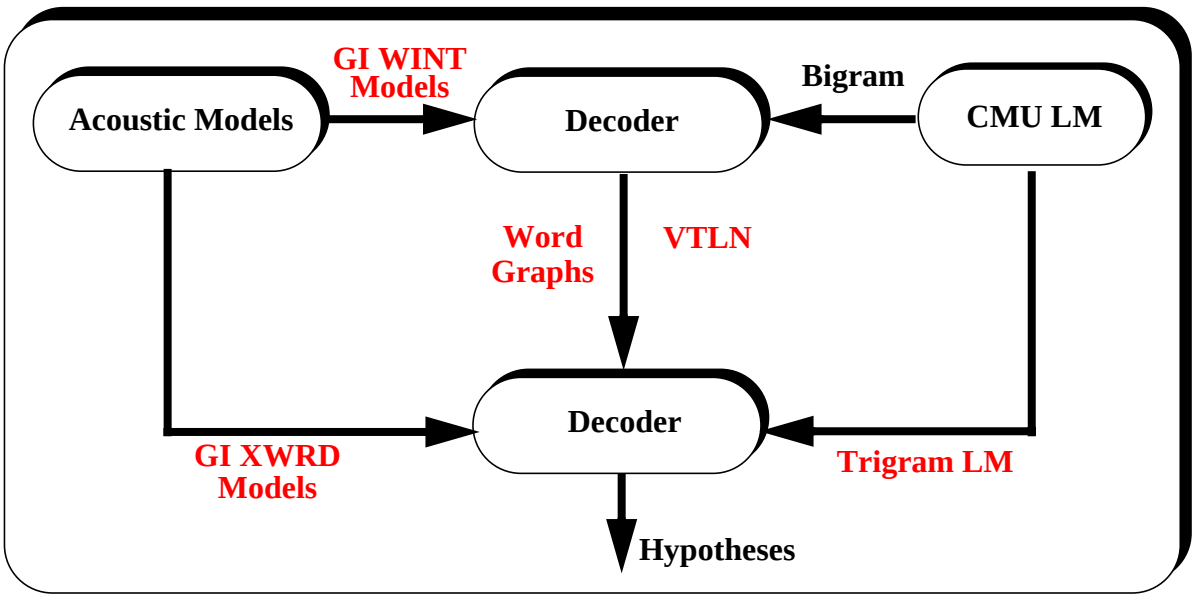


Figure 2. An overview of the decoding procedure used in the SPINE2 evaluation

2.4. SPINE2 Dry-Run Experiments

To minimize computing requirements for a set of fast turnaround experiments, a subset of 300 uncoded and 300 coded utterances was randomly chosen from the SPINE2 dry-run set for testing. The dry-run set originally contained over 4,000 utterances (over one hour of data). Tuning experiments which followed a paradigm suggested by our work in the Aurora project [11] were then conducted on this subset.

Table 1 shows the results of tuning the state-tying parameter for the stage 1 experiments. The goal of these experiments was to achieve the best overall lattice word error rate (WER) on both uncoded and coded speech by optimizing the number of tied states. The optimal number of tied states found in experiment 3 in Table 1 was 1027 (the initial number of states was 2,507). This resulted in a lattice WER of 6.5% and a one-best performance of 38.6% WER.

For stage 2, state-tying tuning was repeated on the above optimized lattices. The optimal number of tied states was 2045, and resulted in a 34.9% WER. This system was our baseline system and the parameters used in this system were used in both training and evaluation.

We also experimented with using VTLN during both training and evaluation. These experiments are summarized in Table 2. We performed an experiment using VTLN only during lattice rescoring, and using VTLN during both lattice generation and lattice rescoring. An interesting observation here is that VTLN did improve the performance of word-internal system but it failed to improve the performance on the final system. The reason is that a 6.5% lattice WER was sufficient for the level of technology evaluated; further improvements in lattice WER were not overcoming the main obstacles to performance at this point.

We achieved a modest reduction in WER on both conditions in Table 2. We can see that there is a 5% absolute improvement in WER using cross-word models. However, both of these enhancements did not result in as much of an improvement on the evaluation set. We believe this

Exp	Tied States	Lattice WER	WINT WER
1	1740	6.5%	38.6%
2	2507	7.3%	44.2%
3	3500	8.3%	44.1%

Table 1. Tuning experiments to optimize the state-tying parameter.

System	WINT Models			XWRD Models		
	un	vo	avg	un	vo	avg
Baseline	32.7	45.7	38.6	27.5	41.3	33.8
VTLN1	32.7	45.7	38.6	25.5	39.9	32.0
VTLN2	30.8	43.6	36.6	25.7	39.4	31.9

Table 2. The impact of VTLN is evaluated on the dry-run subset. VTLN1 is an experiment based on lattice rescoring; VTLN2 was performed on both lattice generation and lattice rescoring. VTLN did improve the performance of the word-internal system, but it failed to improve the overall performance of the final evaluation system.

is one sign that the dry-run set was not a good model of the evaluation set.

In order to compare the performance to last year's SPINE1 system, we took last year's models and decoded on this dry-run subset with two passes of decoding. As shown in Table 3, last year's system achieved a 52.4% WER vs. a 33.8% WER of this year's system (without VTLN). Hence, we made substantial improvements in the acoustic models this year.

2.5. SPINE2 Evaluation Results

The WER's on the final evaluation were 56.9% on uncoded speech and 67.0% on coded speech. A summary of our performance compared to state of the art is shown in Table 4. The best performance this year was a WER of 28.3%, achieved by SRI. The single biggest problem with our system was segmentation — something we specifically disregarded given our limited resources. The degradation due to segmentation was subsequently measured to be 10%. This measurement was based on the reference transcriptions. Some conversation sides had a WER over 100%, indicating significant room for improvement. Many sites' error rates were similarly dominated by these bad conversations. By listening to the data for these conversations we observed that many of these speakers were screaming, yelling and laughing most of the time during conversation.

In order to match our goal of delivering a practical implementation for this task with a limited amount of manpower (one M.S. level grad student working for a few months), we did not put many sophisticated multipass adaptation technologies into this evaluation. However, our result on coded speech seems very promising — only a 14% relative degradation in WER on coded speech as compared to uncoded. Considering that this number was generated from a single set of acoustic models (one system trained to be invariant to this condition), this is an encouraging result.

In Table 5, we provide the error statistics categorized by the type of speech coder. The MELP vocoder is the most promising vocoder, considering it has the lowest bit rate at 2.4 kbps and the lowest relative degradation (9.3%) among the four vocoders. On the contrary, the CVSD vocoder has the highest bit rate at 16 kbps but the relative degradation (29.3%) is triple that of MELP. It seems that the MELP does a good job of preserving the important information in the speech signal. The same result was demonstrated previously in the first quarter of this project using the TIDigits database [12].

System	Lattice WER	WINT WER	XWRD WER
2000	12.5%	54.1%	52.4%
2001	6.5%	38.6%	33.8%

Table 3. A comparison of performance of our 2000 SPINE system to this year's system.

System	Word Error Rate (%)				Best
	Sub	Del	Ins	WER	
Uncoded	28.9	16.5	11.5	56.9	27.5
Vocoded	35.7	18.3	11.1	65.0	53.2

Table 4. The official SPINE2 evaluation results.

2.6. SPINE2 Post-Evaluation Experiments

In order to exclude the effect of poor segmentations, we performed some experiments after the initial evaluation period to diagnose this problem. Table 6 shows the results on uncoded speech. The first experiment was conducted using 45.6% on uncoded speech. Comparing to our 56.9% of evaluation result, we can see that performance improved 11% absolute. This is a substantial improvement in performance. Most sites that worked on utterance segmentation reported that good sentimentalism algorithms only results in a 1% absolute degradation in performance with respect to the reference segmentations. Hence, we feel we could achieve a 10% improvement by simply improving our utterance segmentation.

We ran two additional experiments to calibrate the impact of variance normalization and VTLN. These are summarized in Table 6. First, we retrained models on only uncoded data. Variance normalization [13] was used, which resulted in a 1.3% absolute WER reduction. By performing VTLN, we achieved a 43.7% WER, or a 1.4% relative reduction over the baseline system. VTLN didn't produce as much of an improvement as we expected in this case. VTLN normally gives at least 3% relative reduction in WER, and it did give us 5% relative reduction on the dry-run set. Hence, we doubt the approach we used for VTLN was optimal in this cae. We also observed we had a 12% lattice WER, which was almost double the lattice WER on the dry-run set (6.5%).

The error statistics categorized by the eight noise types are also given in Table 6. To our surprise, the performance on clean data ("quiet" condition) was not the overall best category. It ranked in the middle of the eight noisy conditions in terms of performance and was 10% worse than the "best" performance achieved on the condition "Carrier." This can be explained by noting that the models were biased towards noisy data since training set was dominated by the noisy data (only 1/8 of the data was from the "quiet" condition). Thus, training "quiet" condition dependent models would be a very promising to improve the performance on the "quiet" speech, and probably on the overall evaluation. However, this adds complexity to the system.

As shown in Table 8, we have also conducted a coded speech experiment using the evaluation system and reference segments. A performance of 55.1% WER is very promising comparing to the best performance in this evaluation — 53.2%. We can probably achieve this number by simply performing variance normalization and VTLN. There is only a 9.5% absolute WER degradation from uncoded speech (45.6%) to coded speech, which indicates that our system is fairly robust to coded speech.

Vocoder	Bit Rate (kbps)	Uncoded (%WER)	Vocoded (%WER)	Relative (%)
MELP	2.4	51.8%	56.6%	-9.3%
CELP	4.8	53.0%	60.3%	-13.8%
LPC	2.4	51.3%	59.6%	-16.2%
CVSD	16.0	50.5%	65.3%	-29.3%

Table 5. A comparison of performance for four DoD standard vocoders. MELP coding is extremely robust — it delivers a low bit rate with little degradation in recognition performance.

2.7. Memory and Runtime

For this evaluation, our STT system ran two passes of decoding at

120 xRT on an 800 MHz Pentium processor as shown in Table 9. Since the vocabulary for this application is fairly small by today's standard (5,720 words) and the trigram LM was fairly small, we feel this real-time rate is a bit high. However, we haven't done any serious work on optimizing this system for run time. Recall, however, earlier in this project we demonstrated a real-time Resource Management (RM) system that achieved an 8 xRT baseline system with only a 1% absolute degradation in WER. This is summarized in Table 10. Therefore, it seems highly likely we can make our SPINE2 system run approximately 10 xRT with minimal degradations in performance.

Condition	Eval System				Reference Segmentations			
	Sub	Del	Ins	WER	Sub	Del	Ins	WER
Carrier	20.7	21.2	14.4	56.3	23.7	6.7	8.1	38.5
Car	24.7	19.6	10.5	54.8	27.2	7.8	5.2	40.2
Office	26.1	15.2	8.7	50.0	27.4	8.1	4.7	40.2
Street	29.7	15.2	11.5	56.4	30.8	9.6	6.3	46.7
Quiet	27.8	18.3	13.5	59.6	29.8	10.1	6.9	46.8
Bradley	33.2	13.8	10.8	57.8	34.7	8.1	8.1	50.9
F16	36.9	15.0	11.7	63.6	37.1	9.1	7.9	54.0
Helo	39.5	10.6	11.3	61.4	38.7	7.1	9.6	55.3
Overall	28.9	16.5	11.5	56.9	30.4	8.4	6.8	45.6

Table 6. A comparison of performance for evaluation and a post-evaluation experiment using the reference segmentations on uncoded speech across all the noisy conditions. The substitution errors of recognition are almost equal. However reference segments significantly reduce the deletion and insertion errors.

System (uncoded)	WER
Eval Models	45.6
Variance Normalization	44.3
VTLN	43.7

Table 7. Variance normalization and VTLN produced small gains in performance.

System (coded)	% Error Rate			
	Sub	Del	Ins	WER
Word-internal	37.7	13.3	5.8	56.8
Cross-word	36.8	11.5	6.8	55.1

Table 8. An evaluation conducted on coded speech shows that our system is fairly robust to degradations introduced by compression. The error rate for uncoded speech was 45.6%, and only a 9.5% absolute increase in WER.

Stage	Memory (MBytes)	Runtime (xRT)
Lattice Generation	300	73.8
Lattice Rescoring	600	46.3

Table 9. Run time and memory resources required for the SPINE2 evaluation.

N xRT System	WER	Memory (Mbyte)
8 xRT	3.4%	111
2 xRT	4.3%	57
1 xRT	4.4%	46

Table 10. Performance of a Resource Management system as a function of the real-time rate.

2.8. Hybrid HMM/SVM System

Currently, the dominant approach to acoustic modeling in speech recognition is the Hidden Markov Model (HMM), in which Gaussian mixture models are used to “represent” the various modalities for a given speech sound. The parameters of the Gaussians are estimated using a Maximum Likelihood (ML) criterion. The ML formulation for the representation of the acoustic space does not necessarily translate to better recognition performance since most of the optimization effort is spent in learning the intricacies of the training distributions. Therefore we investigated another powerful classification technique, Support Vector Machines (SVMs), for use in acoustic modeling. The power of SVMs lies in their ability to transform data to a high dimensional space where the data can be separated using a linear hyperplane. This is a highly promising approach for acoustic modeling.

One drawback of SVMs is that they are inherently static classifiers and can not model temporal evaluation of data very well. HMMs have the advantage of being able to handle dynamic data with certain assumptions about stationarity and independence. Taking advantage of the relative strengths of these two classification paradigms we have developed a hybrid SVM/HMM system [14, 15] with HMMs being used to model the temporal information of data and SVMs being used to evaluate acoustic scores. A N-best list based hybrid SVM/HMM system is explained in Figure 3. In this hybrid system, N-best lists with segments are dumped from the conventional HMM system, and then SVM classifiers are used to rescore these segments to generate the final hypotheses.

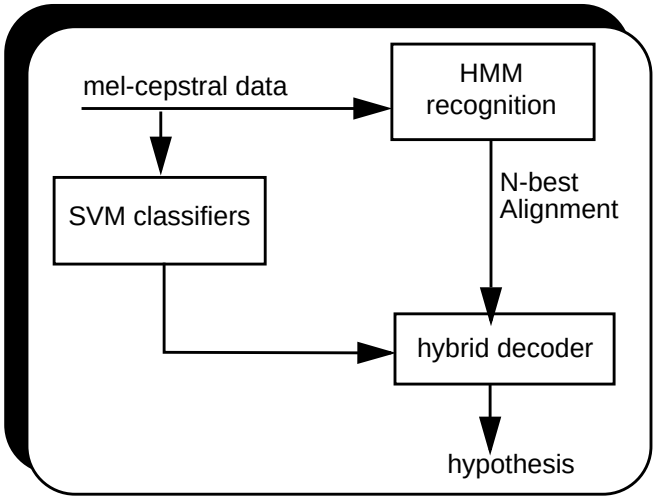


Figure 3. An overview of the hybrid HMM/SVM architecture based on N-best lists.

A preliminary speech experiment was conducted on uncoded data using the hybrid SVM/HMM system and a word-internal HMM system that gave a 32.7% WER. The SPINE2 dry-run subset was used. The hybrid SVM/HMM system rescored N-best lists generated by the HMM system, and achieved a 32.5% WER. Unfortunately, this is not a statistically significant result. However, this SVM system was highly tuned for an alphadigits application [16], and was

limited to using a 10-best list for rescoring. Such a small list is not rich enough for applications such as SPINE2 where performance is uniformly bad. This pilot experiment does suggest, however, that this hybrid SVM/HMM system has a potential for recognition improvement.

3. MULTIPLE-CPU EVAL PACKAGE

In the final quarter of this project, we also delivered a package that allows a user to easily run complete experiments using multiple CPUs. This tool, originally developed for the Aurora project [11], makes it trivial to train and decode in parallel. The user only needs to supply file and transcription lists. The user can specify multiple computers are to be used for training and testing from the command line. The package was modified slightly to allow files with no transcriptions to be decoded. The distribution is available on the project web site [17]. Users must have already installed v5.11 of the ISIP prototype system [18] before running this application. Installation of this package is straightforward using standard Unix configure and make utilities. Detailed instructions are included in the release's AAREADME.text file.

4. N-BEST LIST UTILITY

The ISIP prototype system (v5.11 or earlier) did not officially support N-best hypothesis list generation, though we informally distributed a utility to do this. In order to provide this functionality for our recognition server in the Communicator framework, we had to integrate the N-best list generation utility into our prototype system. This capability will be released with release v5.12 of the prototype system.

We also standardized the format for N-best list output by following a format in use by Lockheed Martin. Here is an example frame that is passed from the recognition server to the Hub:

- {c main
  :session_id "Default"
  :top_nbest "<pause1> cssoc this nine and five over <pause2>"
  :nbest_list ("<pause1> cssoc this nine and five over <pause2>"
  "<pause1> cssoc this ninth and five over <pause2>"
  .....
  "<pause1> cssoc this ninth rat five over <pause2>")
  :server_id {c server_id
  :host "builtin"
  :port -1
  :server_name "Builtin"
  :sockid -2 }
  :utterance_id 0
  :out_lang "english"
  :para_lang "english"
  :synth_lang "english"
  :kv_lang "dialogue"
  :domain "sul"
  :tidx 9 }

The hypothesis string followed by the key word "top_nbest" is the hypothesis with the best likelihood. The "n" strings that follow the key word "nbest_list" are the N-best hypotheses with descending likelihood scores. The remaining information is specific to the Lockheed Martin application, and not modified by the recognizer.

5. FUTURE WORK

Our major milestones this quarter included our participation in the SPINE2 evaluation. We demonstrated a practical triphone/trigram system and achieved a 56.9% WER on uncoded speech and a 67.0% WER on coded speech. Our biggest problem in this evaluation was segmentation. Post-evaluation experiments have been done using reference segments, and the results are promising: a 44.3% WER on uncoded speech and a 55.1% WER on coded speech. A 10% absolute degradation from uncoded speech to coded speech shows our system is fairly robust to coded speech. We also delivered upgrades to our Communicator server, and presented some pilot results on Support Vector Machines [7] at the SPINE2 Workshop.

There are several capabilities that were discussed throughout the course of this project, but were not implemented due to a lack of time. These remaining issues include:

- Instruction of MITRE personnel on how to train a system from scratch using our multiple-CPU eval package;
- Providing a facility to add new words to the system;
- Implementing a dynamic grammar switching capability within the Communicator framework.

Regarding the first item, we delivered a package to MITRE. However, MITRE personnel have not had enough time to work with it and fully understand its capabilities. We plan to support this on an as-needed basis.

The second point is a fairly straightforward piece of technology that involves a mixture of user interface programming and language model smoothing. Similarly, the third point is something we currently have implemented in the released version of the toolkit, but have not incorporated into the Communicator version. The status of future Communicator work at this point is uncertain.

6. ACKNOWLEDGMENTS

We wish to acknowledge Drs. Fred J. Goodman, Bryan George, and George Shuttic of the Signal Processing Center at MITRE Corporation for their continued support and feedback. This one-year project in collaboration with MITRE corporation has been very stimulating and thought provoking. The development of the real-time system enhanced our knowledge about the speech technology and the various parameters associated with the decoder. The Communicator demo that we developed in the course of this project has been an excellent educational tool for novice speech researchers that was demonstrated during our annual design workshop SRSDR’02. The SPINE2 evaluations have helped us to investigate issues involving the effects of a noisy environment and vocoder speech on speech recognition. The knowledge we gained in these efforts will help us improve our technology for robust speech recognition. These opportunities would not have occurred without funding and support from the MITRE Corporation.

7. REFERENCES

[1] J. Huerta, “Speech Recognition in Mobile Environments,” Ph.D. Dissertation, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, April 2000.

[2] “The Second Speech in Noisy Environments Evaluation and Workshop,” *http://elazar.itd.nrl.navy.mil/spine/index.html*, Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington DC, Janaury 2002.

[3] P. Price, W. Fisher, J. Bernstein, and D. Pallett, “The DARPA 1000-Word Resource Management Database for continuous speech recognition,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 651-654, 1988.

[4] L. Hirschman, “Galaxy Communicator,” *http://communicator.sourceforge.net*, MITRE Corporation, Bedford, Massachusetts, USA, April 20, 2001.

[5] “Evaluation of Speech Technology in Noisy Environments Workshop,” *http://elazar.itd.nrl.navy.mil/spine/spine1/index.html*, Naval Research Laboratory, Washington, DC, October 2000.

[6] B. George, B. Necioglu, J. Picone, G. Shuttic, and R. Sundaram, “The 2000 NRL Evaluation for Recognition of Speech in Noisy Environments,” presented at the Speech In Noisy Environments (SPINE) Workshop, Naval Research Laboratory, Alexandria, Virginia, USA, October 2000.

[7] F. Zheng, F. Goodman, J. Hamaker, B. George, N. Parihar, and J. Picone, “The ISIP 2001 NRL Evaluation for Recognition of Speech in Noisy Environments,” presented at the Speech In Noisy Environments (SPINE) Workshop, Orlando, Florida, USA, November 2001.

[8] “The U.S. Department of Defense Digital Voice Processor Consortium,” *http://www.plh.af.mil/ddvpc/index.html*, Department of Defense, USA, January 1998.

[9] N. Deshmukh, A. Ganapathiraju, J. Hamaker, J. Picone and M. Ordowski, “A Public Domain Speech-to-Text System,” *Proceedings of the 6th European Conference on Speech Communication and Technology*, vol. 5, pp. 2127-2130, Budapest, Hungary, September 1999.

[10] R. Singh, “CMU Language Model for SPINE2 Evaluation,” *http://www.cs.cmu.edu/~rsingh/spine2lm.tar.gz*, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, September 2001.

[11] N. Parihar, “Aurora Evaluations,” *http://www.isip.msstate.edu/projects/aurora/index.html*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2002.

[12] F. Zheng and J. Picone, “Robust Low Perplexity Voice Interfaces,” *MITRE Corporation*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, May 15, 2001.

[13] T. Hain, P.C. Woodland, T.R. Niesler, and E.W.D. Whittaker, “The 1998 HTK System for Transcription of Conversational Telephone Speech,” *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, USA, pp. 57-60, March 1999.

[14] A. Ganapathiraju, J. Hamaker and J. Picone, “Hybrid HMM/SVM Architectures for Speech Recognition,” *Proceedings of the Speech Transcription Workshop*, College Park, Maryland, USA, May 2000.

[15] A. Ganapathiraju, J. Hamaker and J. Picone, “A Hybrid ASR System Using Support Vector Machines,” *Proceedings of the International Conference of Spoken Language Processing*, vol. 4, pp. 504-507, Beijing, China, October 2000.

[16] “Alphadigit V1.1,” *http://cslu.cse.ogi.edu/corpora/alphadigit/*, Center for Spoken Language Understanding, OGI School of Science and Engineering at OHSU, Beaverton, Origon, USA, January 2002.

[17] F. Zheng, “Robust Low Perplexity Voice Interfaces,” *http://www.isip.msstate.edu/projects/robust_low_perplexity/index.html*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2002.

[18] N. Parihar, “ASR Downloads,” *http://www.isip.msstate.edu/projects/speech/software/downloads/*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2002.