Building an ASR System for Noisy Environments: SRI's 2001 SPINE Evaluation System

Venkata Ramana Rao Gadde, Andreas Stolcke, Dimitra Vergyri, Jing Zheng,
Kemal Sonmez, and Anand Venkataraman

Speech Technology and Research Laboratory, SRI International
333 Ravenswood Avenue, Menlo Park, CA 94025, USA.
http://www.speech.sri.com/

ABSTRACT

We describe SRI's recognition system as used in the 2001 DARPA Speech in Noisy Environments (SPINE) evaluation. The SPINE task involves recognition of speech in simulated military environments. The task had some unique challenges, including segmentation of foreground speech from noisy background, the need for robust acoustic models to handle noisy speech, and development of language models from limited training data. In developing the SRI evaluation system for this task, we addressed each of these challenges using a combination of state−of−the−art techniques, including several types of feature normalization, model adaptation, class−based language modeling, multi−pass segmentation and recognition, and word posterior−based decoding and system combination

1. INTRODUCTION

We describe SRI's evaluation system for the October 2001 Speech in Noisy Environments (SPINE) task. The main aim of the paper is to present the key algorithms and components of a state−of−art speech recognition system and how they were combined into a system for optimal performance.

We have organized the paper as follows. First, we provide a brief introduction to the SPINE task and its challenges. We then present the key components and algorithms used in our system and how the task features guided the design of the system. We then present the results on two test sets, the dry run and the evaluation sets of the 2001 SPINE evaluation. We conclude with a discussion of the results.

SPINE is a relatively new task developed by the Naval Research Laboratory (NRL) to test the state of the art of speech recognition in military noise environments. The primary challenge of the task is recognition of speech with significant amounts of background noise as found in various military environments, such as fighter jet cockpits and aircraft carrier flightdecks. The data consists of dialogs between two participants playing a battleship−like game, with recorded military noises played back into the recording environment. The players use realistic microphones and headgear (e.g., fighter pilot helmets) as appropriate for the different scenarios. The language used comprises a mix of commands, status reports, and confirmations specific to this limited domain, involving an active vocabulary of about 2000 words. More details are available at [1].

Due to its focus on noisy environments, SPINE posed some unique challenges. One of the difffculties was to segment foreground speech from the noisy background, which, in some environments, included background speech. Another challenge was to develop robust acoustic features, models, and techniques capable of recognizing the noise−degraded speech. Yet another challenge was the limited amount of training data, particluarly for training the language model. Thus, the task posed challenges not only for research, but also for system development. To solve these issues of segmentation and robust acoustic and language modeling, we drew on a number of state−of−the−art algorithms as the building blocks of our system. In the following, we describe these components and how they were integrated into a system that achieved the lowest word error rate (WER) in the 2001 SPINE evaluation.

2. SYSTEM DESCRIPTION

2.1 Key Components and techniques

Various components and techniques were used to achieve robustness.

2.1.1. Robust segmentation

As mentioned above, one of the unique challenges in SPINE is the problem of segmentation. Since the data included foreground speech with significant amounts of background noise and speech, traditional segmentation approaches that rely primarily on energy performed poorly. In the 2000 SPINE evaluation, bad segmentation was the main reason for poor performance across all competing sites [1]. To solve the segmentation problem, we used the following four−stage approach:

1. Classify the waveform into foreground (speech) and background (noise, low amplitude speech, silence) using a 2−class hidden markov model (HMM), trained with forced alignments of the acoustic training data. Apply constraints on segment duration to merge very short segments with neighbors.
2. Recognize the foreground segments using gender− and speaker−independent acoustic models and the SPINE language model. Generate N−best lists of hypotheses.
3. Combine the N−best lists into a confusion network and estimate the posterior probabilities for all hypothesized words [2].
4. Resegment foreground segments, eliminating regions corresponding to words with low posteriors.

The goal of the last step is to further eliminate

background speech and noise; foreground words recognized with low posteriors will also be eliminated, but would in all likelihood have been misrecognized anyway. Our segmentation algorithm resulted in recognition performance close to that obtained with segmentations produced by human transcribers.

2.1.2. Robust acoustic modeling

An important challenge in SPINE is to develop robust acoustic features and models to handle noisy speech. We used a multipronged approach to achieve this. First, we used features normalized for speaker and environment variations. Second, we used the normalized features to build acoustic models and applied model adaptation techniques to further reduce the mismatch between the acoustic models and the evaluation data. Third, we used multiple features and built multiple recognition systems based on these, and combined their outputs using a posterior−weighted version of the ROVER algorithm [3,4]. This approach led to significant improvements in the system performance as shown later. Here we describe each of these three approaches in more detail.

Acoustic features. To improve the robustness of the system, and inspired by CMU's SPINE1 system [5], we built three different systems based on three different features. The features were derived from the Mel frequency cepstrum (MFC), the perceptual linear prediction (PLP) cepstrum [6], and the linear frequency cepstrum (LFC) [7], respectively. Each feature vector consisted of 13 cepstra, 13 derivatives and 13 second derivatives. We then normalized the features, using vocal tract length normalization [8], to minimize mismatch between the test speakers and the models. We also performed feature mean and variance normalization to reduce channel mismatch. Finally, we normalized the features, using a linear transformation, estimated with the constrained maximum likelihood linear regression (MLLR) algorithm [9], to further reduce speaker and environment mismatch. These normalized features were both used during model training and testing.

Acoustic modeling. We trained speaker− and gender− independent acoustic models (genonic HMMs [10]), using normalized features. To build robust models, we pooled clean and noisy training data. The models were trained using several iterations of the Baum−Welch algorithm, followed by one additional iteration of discriminative training using the MMIE algorithm [11]. For each feature, we trained models for both crossword and noncrossword triphones.

Acoustic adaptation. We used adaptation as the primary tool to compensate for noise, using two different techniques. First, for each speaker, we estimated a global adaptation transform using the constrained MLLR algorithm. The inverse of this transformation was applied on the input features. This was done in training of acoustic models and during evaluation, in a way similar to feature−based speaker adaptive training [12].

In addition to the feature transformation, we adapted model means and variances using a modified MLLR algorithm. We used seven separate transforms for phonetically motivated phone classes to improve performance.

2.1.3. Robust language modeling

The recognition vocabulary consisted of about 5,200 words, the same as that of the reference language model provided by CMU to all evaluation participants. This vocabulary included the most frequent words from the Switchboard corpus, in addition to those found in the SPINE training corpus. We augmented this vocabulary with about 800 multiwords as required by our pronunciation models [4].

Since SPINE is a task−oriented corpus, and since only a relatively small amount of training data was available, we chose a word−class−based N−gram model to improve generalization. The main relevant task characteristic was that a set of predefined words, the "grid voculary", was used to refer to coordinates in the battleship−like game. Furthermore, the grid vocabulary had been changed after the first SPINE data collection, and was therefore incompatible between a portion of the training material and the test set. To complicate matters further, some of the grid words were ordinary English words and thus ambiguous between their grid word usage and their general meaning.

We developed an HMM−based tagger to identify and disambiguate grid words in the training corpus. The tagger was trained in unsupervised fashion, bootstrapping from grid words that were unambiguous. One word class was dedicated to all grid words, with uniform membership probabilities. An additional word class was created for spelled grid words, which are easily identified by matching spelled letter sequences to the vocabulary. All other words were put in singleton word classes.

A further refinement of the language model (LM) was achieved by training separate LMs from the SPINE1 and SPINE2 phases of the training corpus, and interpolating them with a fixed weight that was tuned on the development data. Thus, we created interpolated class−trigram and 4−gram models for use in decoding and N−best rescoring.

2.1.4. Word posterior−based decoding and system combination

The speaker−adapted acoustic models for the three features (MFC, PLP, LFC) were used with a trigram LM to produce as many as 2000−best hypotheses per waveform segment. These were then rescored with additional knowledge sources: the pronunciation probabilities and the class 4−gram LM. The hypotheses for each system were aligned into word confusion "sausages", and posterior probabilities were computed for each word [2,4]. To obtain the best overall hypothesis, the highest posterior word at each position in the alignments was chosen. The score weights for each of the knowledge sources (acoustic model, LM, pronunciation model, and insertion penalty), and the posterior scaling factor, were jointly optimized for each system to minimize the number of word errors.

To combine multiple systems, all N−best lists for one utterance were aligned into a single confusion network ("N−best ROVER"). The weights and scaling factors optimized as described were used to obtain word posteriors for each system, and the total word posterior of the system combination was computed as the weighted average of system−specific posteriors (weighted equally).

2.1.5. Acoustic model readaptation

We took advantage of system combination to iteratively improve the hypotheses used in acoustic model adaptation, similar to [5]. However, we did not observe incremental accuracy improvements beyond the first iteration, and even then

only when taking special precautions to prevent the three combined systems from converging as a result of the shared adaptation hypotheses. To maintain model diversity prior to the final system combination, we used a "leave−one−out" technique: each of the three acoustic models (MFC, PLP, LFC) was adapted to the combined output of the *other* two systems.

2.1.6. Processing steps

Combining all these techniques, the recognition system proceeds in the following steps:

1. Segment the conversation−length waveforms.
2. Compute vocal tract length and cepstral means and variances per conversation side.
3. Do first pass recognition using gender− and speaker− independent noncrossword acoustic models and trigram language model.
4. Use the hypotheses from Step 3 to compute feature transformations for each speaker (for all three features).
5. Perform a second recognition pass using the transformed features (for all three features).
6. Perform unsupervised transcription−based adaptation of the means and variances of the noncrossword and crossword models to each speaker, using the hypotheses from Step 5 (for all three features).
7. Use the speaker−adapted noncrossword acoustic models from Step 6 to generate lattices (for all three features).
8. Recognize from lattices using the speaker−adapted crossword models and generate 2000−best list. Rescore the N−best hypotheses using the 4−gram language model and pronunciation probabilities (for all three features).
9. Readapt all three models to the result of pairwise combinations of the three systems.
10. Repeat Step 8.
11. Combine hypotheses from all three systems using N−best ROVER.

3. EXPERIMENTS

We report the results of experiments conducted using two test sets, the SPINE 2001 dry run set and the evaluation set. The results are shown only for the MFC system, though similar results were observed for all systems.

3.1. Segmentation

Table 1 shows the recognition WER when we used our segmenter and compares it with NRL−supplied manual segments. The increase in WER due to automatic segmentation is only about 1−2% relative.

Test Set	NRL (manual) segmentation	SRI (automatic) segmentation
Dry run 2001	31.3	31.6
Eval 2001	38.2	39.0

Table 1. WERs for different segmentations.

3.2. Acoustic adaptation

Adaptation served as our primary tool to build robust acoustic models. Two types of adaptation were used in our system. In Table 2 we show the reduction in WER obtained from feature transformation based on the constrained MLLR algorithm.

Test Set	Untransformed	Transformed
Dry run 2001	31.6	27.1
Eval 2001	39.0	34.9

Table 2. Effect of feature transforms on WER.

Table 3 shows the improvements from model adaptation using a modified MLLR algorithm incorporating model mean transforms and variance scaling.

Test Set	Unadapted	Adapted
Dry run 2001	27.1	24.5
Eval 2001	34.9	33.0

Table 3. Effect of model adaptation on WER.

The results show that combined feature and model transform−based speaker and environment adaptation reduces WER significantly, by 22% on the dry run set and 15% on the evaluation set.

3.3. Language model performance

Table 4 compares different types of trigram LMs in terms of both perplexity and WER, decoding the dry run test set in the first recognition pass. The baseline model was the reference LM supplied by CMU. "Interpolated" LMs were obtained by training separate models from SPINE1 and SPINE2 training data, and mixing their probabilities at the word level, as discussed above. The "naïve" class LM was obtained by replacing every potential grid word by its class label in training, whereas the "tagged" class LM used the unsupervised HMM tagger to disambiguate such occurrences. The results show that both model interpolation and tagger−mediated class labeling improve LM performance.

Model	Type	Perplex.	WER
Baseline trigram	Word	58.6	36.9
Interpolated	Word	50.9	−
Interpolated	Class, naïve	43.7	31.7
Interpolated	Class, tagged	39.7	31.2

Table 4. Language model performance comparison. The word−based interpolated LM was not evaluated in recognition.

3.4. System combination

Table 5 shows the results of system combination at three steps. The first three rows show the WERs for the three systems, when the N−best hypotheses were generated (at Step 8). The following row shows the results of combining the three systems after processing Step 8. The last row contains the results of the final system combination after readaptation (Step 11).

We see that system combination achieves a relative WER reduction of 15% on the dry run data, and 11% on the evaluation data, compared to the best of the individual systems.

Acoustic model readaptation, however, contributes only a very small WER reduction.

WER		Test Set	
		Dry run 2001	Eval 2001
Individual systems	MFC	23.5	32.5
	PLP	23.2	31.5
	LFC	22.7	31.9
1st combined system (Step 8)		19.5	28.1
2nd combined system (Step 11)		19.3	28.0

Table 5. Results of multiple system combination.

3.5. Other techniques

In addition to the significant reductions in WER shown above, we obtained smaller reductions from a few other techniques.

- A dialog language model showed inconsistent performance, giving 0.2% reduction on the dry run data, but no improvement on the evaluation data.
- In the final recognition output we omitted words with posteriors below a tuned threshold, since such words most likely represent noise−induced insertions or misrecognitions. This further reduced the WER by 0.1%.

4. CONCLUSIONS

We have presented the key components and techniques used in the SRI 2001 SPINE evaluation system. A new segmenter, combining a 2−class HMM recognition stage with posterior−based word rejection, minimizes recognition degradion due to segmentation error to less than 1% absolute. Other techniques include the 2−stage adaptation for both features and models and word posterior−based decoding, readaptation and combination of recognition systems using MFC, PLP, and LFC front ends. While many of these techniques are well known, their careful combination is unusually effective in the context of the SPINE task. Furthermore, more recent experiments show that a similar system architecture is equally effective for other recognition tasks, such as large vocabulary conversational ASR in the Switchboard domain.

The SPINE language model benefited greatly from knowledge of the task, and in particular from task−specific word classes. The challenge for SPINE and other tasks will be to automatically induce task−given structure for LM purposes, for example, in the dialog, without requiring human guidance or annotation.

Other areas for ongoing research in the SPINE domain include the use of duration models (which have been successful in spontaneous speech [13]) and prosodic information in general.

5. ACKNOWLEDGMENTS

We thank Dr. Dan Ellis and the International Computer Science Institute (ICSI) for providing the implementation of the PLP features used in our system. This research was in part funded by the DARPA ROAR program under contract N66001−99−D−8504, D.O. 0003. The views herein are those of the authors and should not be interpreted as representing the policies of the funding agency.

6. REFERENCES

[1] Navy Research Laboratory, SPINE2: Speech in Noisy Environments, http://elazar.itd.nrl.navy.mil/spine/,2001.
[2] Mangu, L., Brill, E., and Stolcke, A., Finding Consensus in Speech Recognition: Word Error Minimization and other Applications of Confusion Networks, Computer Speech and Language 14(4), 373−400, 2000.
[3] Fiscus, J.G., A Post−processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). Proc. IEEE Automatic Speech Recognition and Understanding Workshop, pp. 347−352, Santa Barbara, CA, 1997.
[4] Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Gadde, V.R.R., Plauche, M., Richey, C., Shriberg, E., Sonmez, K.M., Weng, F., and, Zheng, J., The SRI March 2000 Hub−5 Conversational Speech Transcription System, Proc. NIST Speech Transcription Workshop, College Park, MD, 2000.
[5] Singh, R., Seltzer, M.L., Raj, B., and Stern, R.M., Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination, Proc. IEEE ICASSP, vol. 1, pp. 273−276, Salt lake city, 2001.
[6] Hermansky, H., Perceptual Linear Predictive (PLP) Analysis of Speech, Journal of the Acoustic Society of America 87(4), 1738−1752, 1990.
[7] Gadde, V.R.R, Linear Frequency Cepstral Features for Speech Recognition, submitted to ICSLP 2002.
[8] Wegmann, S., McAllaster, D., Orloff, J., and Peskin, B., Speaker Normalization on Conversational Telephone Speech, Proc. IEEE ICASSP, vol.1, pp.339−341, 1996.
[9] Gales, M.J.F., Maximum Likelihood Linear Transformations for HMM−based Speech Recognition, Computer Speech and Language 12(1), 75−98, 1998.
[10] Digalakis, V., and Murveit, H., GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model based Speech Recognizer, Proc. IEEE ICASSP, vol. I, pp. 537−540, 1994.
[11] Zheng, J., Butzberger, J., Franco, H., and Stolcke, A., Improved Maximum Mutual Information Estimation Training of Continuous Density HMMs, Proc. EUROSPEECH, vol. 2, pp. 679−682, Aalborg, Denmark, 2001.
[12] Jin, H., Matsoukas, S., Schwartz, R., and Kubala, F., Fast Robust Inverse Transform SAT and Multi−stage Adaptation, Proc. DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, pp. 105−109, 1998.
[13] Gadde, V.R.R, Modeling Word Duration for Better Speech Recognition, Proc. NIST Speech Transcription Workshop, College Park, MD, 2000.