The SRI Spine 2000 Evaluation System

Venkata Ramana Rao Gadde
Andreas Stolcke

Speech Technology And Research Laboratory
SRI International

Organization of the Talk

- The Spine 2000 task
- SRI's Evaluation System
- Post-evaluation improvements
- Future work

The Spine 2000 task

- Evaluation of current ASR technology in noisy military environments.
- Differences from the Hub-5 task
  - Evaluation data is not segmented.
  - All sites must use a common language model.
  - Sites are not funded.

SRI' Evaluation System

0. Segmentation of speech.
1. Cluster segments and estimate front-end normalizations (VTL, cepstral mean and variance).
2. First pass recognition using SI acoustic models and 3-gram multiword language model.
3. Adapt the SI acoustic models to the clusters.
4. Dump Nbest (N=2000) using the cluster adapted acoustic models and the 3-gram multiword language model.
5. Rescore the Nbest using 3-gram language model (non-multiword).
6. Format the results for submission.

Models

- Acoustic Models
  - Acoustic models trained from the Spine training data (11970 waveforms). No DRT data was used.
  - Clustering was used to identify 'pseudo speakers'
- Language Models
  - Two language models were used, the CMU 3-gram LM and a 3-gram multiword LM derived from the CMU LM.

Language Model

Need to convert CMU language model to contain multi-word units used in dictionary.

- Insert all multi-word N-grams that are "triggered" by original N-grams. Example:

  Old N-grams: `i'm going`, `to do`
  Multiword: `going_to`
  Add N-grams: `i'm going_to`, `going_to do`
  Remove N-gram: `going to`

- Assign probabilities so that word sequences retain combined probabilities:

$$p(\texttt{going\_to}|\texttt{i'm}) = p(\texttt{going}|\texttt{i'm}) \times p(\texttt{to}|\texttt{i'm going})$$

Front-end processing

- Segmentation

  0. Split the two channel waveform by conversation side.

  1. Remove 'digital zeros' from the waveforms.

  2. Recognize the waveforms using gender and speaker independent acoustic models and a multiword bigram language model. Use the recognition hypotheses to further segment the waveforms into speech/nonspeech.

  3. Perform foreground/background speech classification using energy. Use it to obtain foreground speech segments.

- Clustering

  0. Cluster the foreground speech segments using a bottom up agglomerative clustering scheme (SRI's 1997 Hub4 eval system).

  1. Compute cluster level normalizations (VTL, cep. mean and var.).

N-best Rescoring

- Replace multi-words with component words
- Recompute language model probabilities using CMU LM

  Note: This yields better results because step 1 in multiword LM construction gives only an approximation to full multi-word N-gram.
- Align all N-best hypotheses and extract words with highest posterior probabilities (explicit word error minimization; Stolcke, Konig, & Weintraub 1997)

Results on development set

- Two dev sets were taken out of the training set, one containing speech from selected speakers and a second containing all the _nv_ data.
- Acoustic models were trained from the remaining data.
- LM for each set was trained from the transcripts, excluding the transcripts for that set.

Model	Test set		
	Set 1	set 2	both
step 2. Rec with SI	35.0%	41.1%	36.8%
step 4. Rec with Adapted	32.8%	39.1%	34.7%
step 5. Rescore Nbest	32.1%	37.1%	33.9%

- The improvements are similar to our Hub-5 system.
- Using a larger bandwidth front-end gave a small reduction in WER (not used in our eval system).
- Clustering the training data was better than using speaker/noise labels.
- Multiwords gave 1-2% improvement in WER.
- Using probabilities for pronunciations gave a small reduction in WER.

Evaluation results

- SRI's evaluation system had a WER of 46.3%. The best system had a WER around 26%.
- Reasons for the poor performance
  - Incorrect segmentation
    - Large number of insertions and deletions.
    - Loss of 12.5% absolute due to incorrect segmentation.
    - Could not tune segmentation thresholds
      - lack of representatitive dev data
      - missed early clarification on what to do with background speech
  - Simpler system compared to our Hub-5 system
    - No crossword models
    - No duration models
    - No rate-of-speech models

Post-evaluation Improvements

- Improvements in segmentation
  - Segmentation algorithm is simplified, using only energy.
  - Thresholds for segmentation optimized using the dev set.
  - WER reduced by 7.5%
- Using Crossword models
  - Crossword acoustic models were used to rescore the lattices.
  - WER reduced by 6.0% absolute on dev set.

Post-evaluation System

0. Segmentation of speech.
1. Cluster segments and estimate front-end normalizations (VTL, cepstral mean and variance).
2. First pass recognition using SI acoustic models and 3-gram multiword language model.
3. Adapt the SI acoustic models to the clusters.
4. Adapt the crossword SI acoustic models to the clusters.
5. Generate lattices using the adapted models and a bigram multiword LM. Expand using the 3-gram LM.
6. Dump Nbest (N=2000) from lattices using the cluster adapted crossword acoustic models.
7. Rescore the Nbest using 3-gram language model (non-multiword).
8. Format the results for submission.

Comparison of the Eval and Post-eval Systems

Step	Model	
	Eval	Post-eval
step 2. Rec with SI	52.2%	42.9%
step 3. Rec with Adapted	49.5%	—
step 6. Rec with Adapted CW	—	36.3%
step 7. Rescore Nbest	48.5%	35.8%
NIST scoring	46.3%	33.1%*

* - projected value

Future Work

- Segmentation
  - Foreground/background speaker classification
  - Prosody based segmentation
- Acoustic modeling
  - Utilize noise characteristics
  - Spectral subtraction did not help
- Duration modeling

Suggestions for Future Evaluations

- Need clearly defined training and dev sets.
  - In absence of a dev set, we were unaware of background speech issue till a week before the evaluation.
- Common LM was an unnecessary constraint. Sites should be allowed to build their own LMs using a common training data.
  - We could try to model dialog instead of sentences.
- Noise recordings (used in preparing the data) could be made available.