The SRI RT-02 Speech-to-Text System

A. Stolcke R. Gadde

A. Venkataraman D. Vergyri J. Zheng

Speech Technology and Research Laboratory

SRI International

C. Wooters

International Computer Science Institute

Overview

- System architecture overview
- Evaluation result breakdowns
- Improvements over last year's system
  - Model-based feature normalization (SAT)
  - System combination with PLP and LFC features
  - Posterior smoothing in training and adaptation
  - Improved adaptation of MMIE models
  - Pause language model
  - Miscellaneous other improvements
- Hidden event prosody model (D. Vergyri)
- Meeting recognition (C. Wooters)

Basic System Overview (1)

1. Gender ID using 2-class GMMs
2. VTL estimation and feature normalization
3. Adapt within-word models using phone-loop
4. N-best w/adapted within-word models
5. Rescore N-best with
   - interpolated class 4-gram LM
   - phone-in-word duration model (Gadde 1999, 2000)
   - pause LM
   - apply word posterior maximization (sausage decoding)
6. Estimate speaker-specific feature transforms
7. Adapt SAT models to Step 5 hypotheses
8. Make bigram lattices using within-word SAT models

Basic System Overview (2)

9. Expand lattices using trigram LM
10. N-best w/cross-word adapted SAT models
11. Rescore N-best with
    - interpolated class 4-gram LM
    - phone-in-word duration model (Gadde 1999, 2000)
    - pause LM
    - apply word posterior maximization (sausage decoding)
12. Confidence estimation using multi-layer perceptron (warping of word posteriors)

This describes the submitted sri2 contrast system.

Front end used Mel frequency cepstrum (MFC).

SRI International

Full System Overview

Primary evaluation system (sri1) had additional steps:

- Steps 2-6 and 10-11 were repeated with alternate front ends: PLP and linear (Fourier) frequency cepstrum (LFC)
- All 3 systems shared lattices generated by MFC system.
- Step 11 outputs were combined pairwise to re-adapt the third system, round-robin.
- Steps 7, 10, 11 repeated using readapted models.
- 3 final system outputs were combined with N-best ROVER (word error minimization using interpolated word posterior estimates).

Acoustic Model Diversity

- System combination of models of different types is effective (cf. 2001 system)
- MFC, PLP and LFC cross-word acoustic models also differed in type:
  - MFC models used rate-dependent phone sets (Zheng et al. 2000)
  - PLP models used MMIE training
  - LFC models used standard phone set and MLE training
- Rate-dependent training and MMIE give independent improvements.

Switchboard Evaluation Results

Step	2002	2001 eval set	
	2002	2002/2001 system	
4. PL-adapted 1-best	40.2	37.4	39.0
5. rescoring	35.8	33.1	34.8
8. lattice generation	35.5	33.1	36.0
10. CW-adapted 1-best	31.7	29.7	33.4
11. rescoring	29.4	27.5	30.6
12. CW-readapted 1-best	30.2	28.0	32.7
13. rescoring	28.2	26.0	29.9
14. system combination	27.4	25.3	29.0

Notes: Steps 11-13 in 2002 system based on PLP system.

Switchboard Results: Comments

- Development and tuning done using subset of official 2001 development set (*not* on eval2001).
- 2002 eval set was about 2% (absolute) harder than 2001 set.
- All incremental improvements almost identical on dev2001, eval2001, and eval2002.
- Improvement over 2001 system:
  3.7% absolute, 13% relative (on eval2001)

SRI International

Meeting Recognition System (1)

Personal microphone, PE condition

- Identical to basic Switchboard system
- Waveforms downsampled to 8000 kHz
- No meeting data used for model training!
- Gender ID, feature normalization and phone-loop adaptation on full meetings
- Recognition and later passes run only on 10-minute evaluation segments
- Development very difficult due to
  - Lack of data and time
  - Only 5 transcribed training meetings available
  - No NIST meeting samples

Meeting Evaluation Results (1)

Personal mics, UE condition

	ALL	icsi	cmu	ldc	nist	SWB
PL-adapted 1-best	44.9	34.6	55.6	46.7	44.1	40.2
rescored	41.5	30.6	54.0	43.0	39.9	35.8
SAT lattice gener.	40.6	30.6	51.9	42.5	39.9	35.5
CW-adapted 1-best	37.8	27.8	49.7	38.7	36.6	32.6
rescored	36.0	25.9	47.9	36.8	35.2	30.2

- ICSI meeting easiest, CMU hardest (also on devtest data)
- Incremental improvement comparable to SWB
- Further evidence that SWB is “ASR-complete”

Meeting Recognition System (2)

Tabletop microphone, UE condition

- Additional preprocessing
  - Filter waveforms with noise-reducing filter from Qualcomm-ICSI-OGI Aurora-2 system
  - Detect speech and segment using 2-class HMM
  - Cluster segments into 5 pseudo-speakers for normalization & adaptation (similar to SRI Hub-4 system)
  - Used full meetings in preprocessing
- Skip lattice generation & expansion
- Skip 2nd recognition pass on CMU meetings after bad performance on devtest data.
- Final rescoring used bigram LM.
- Segmenter trained on 1 ICSI + 1 CMU meeting only

Meeting Evaluation Results (2)

Tabletop mic, UE condition

	ALL	icsi	cmu	ldc	nist
PL-adapted 1-best	65.8	57.8	69.8	72.4	66.9
rescored	63.1	55.3	66.8	70.6	62.9
SAT-adapted 1-best	61.9	53.7	65.1	69.6	62.8
rescored	61.6	53.6	64.5	69.7	61.6

- 2nd pass on CMU meetings was run post-evaluation.
- Rescoring and adaptation less effective than with personal mic, PE system.
- WER about double of personal microphone, UE system.

SRI International

Improvements in LVCSR

Model-based Feature Normalization

- Implemented inverse-transform SAT (Jin et al. 1998)
- Estimate feature transforms using ML w.r.t. 1st pass recognition models.
- Iterated model estimation did not give improvements.
- Normalize training speakers using transcripts; test speakers using 1st pass rescored hyps
- MLLR operates on SAT-normalized features
- Use full transforms in training, block-diagonal in testing

	WER (devel2001)
MLLR w/o SAT	35.1
MLLR with SAT, full transforms	34.0
block-diagonal transforms	33.7

3-Front End Combination

- Combined MFC, PLP, and LFC systems using N-best ROVER word posterior maximization
- 2-way leave-one-out ROVER used to re-adapt acoustic models (Hub-5 2001)
- LFC shows much higher WER in early passes, less so in later passes (cf. SPINE results)

	WER (dev2001)			
	MFC	PLP	LFC	Combined
PL-adapted	33.6	34.2	39.9	32.3
CW adapted	28.9	28.5	30.2	27.1
CW re-adapted	27.2	27.2	28.0	26.4

Forward-backward Posterior Smoothing

- Posterior-based decoding shows: posteriors are best estimated by dividing acoustic likelihoods by language model weight (Stolcke et al. 1997)
- Acoustic score scaling leads to smoother state posteriors in forward-backward algorithm
- Improves MMIE training (Woodland & Povey 2000)
- Also helps MLE acoustic model training & adaptation!

	WER (dev2001)
No smoothing	31.3
Smoothing in adaptation only	31.2
Smoothing in training & adaptation	30.9

Adaptation of MMIE-trained Models

- Does MLE in adaptation undo discriminative model estimation in training?
- Idea: estimate adaptation transforms using MLE-trained (numerator) models, then apply them to MMIE-trained models.

	WER (male dev2001)
Unsmoothed MLE models	30.7
Smoothed MLE models	30.3
MMIE models in adaptation	29.9
MLE models in adaptation, & MMIE models in recognition	29.6

Pause Language Modeling

- Observation: pauses are accounted for in acoustic models, but not in LM.
- (Optional pauses have no penalty in search, ignored in rescoring.)
- Including pauses in standard LM would fragment N-gram space.
- Idea: model pauses as separate knowledge source, conditioned on left and right word (trigram model)
- Total utterance likelihood is decomposed as

  P(W) x P(pauses | W) x P(acoustics | W, pauses)

Pause LM: Results

- Trigram backoff pause LM:

  P(pause-type | left-word, right-word)
- 3 pause types:
  - pause $\geqslant$ 0.6 sec
  - 0.6 sec > pause $\geqslant$ 0.06 sec
  - pause < 0.06 sec
- Complements phone-in-word duration model

	WER (dev2001)
Baseline (incl. duration model)	31.4
With pause LM	31.0

Other Improvements

- Improved phone-in-word duration model (0.3% abs.)
  - Retrained on full SWB training set
  - Model off-diagonal covariances of phone durations
  - Ignore noise and pause models
- More phone-classes in adaptation (0.1-0.3% abs.)
  - Increased number of classes from 7 to 10 and 8, for PLP and LFC features, respectively.
- More and cleaner training data (1.0% abs.)
  - Supplemented old BBN training set with MSU-ISIP data.
  - Recognized all training speakers and eliminated those with high WERs.
  - High WERs due to bad transcripts, noise, or crosstalk.

Our To-Do List

SRI STT system is currently still lacking the following commonly found features:

- LDA/HLDA feature transforms
- phone context beyond triphones
- cross-word context-dependent phones and trigrams in first decoding

Lots of room for improvement!

Hidden Event Modeling Using Language and Prosody

D. Vergyri L. Ferrer
E. Shriberg A. Stolcke

SRI International

Prosody for Improved LVCSR

- Prosodic features of speech (suprasegmental duration, pause, pitch, and energy) have been found useful for various tasks but are still not used in the state-of-the-art ASR systems.
- Prosody correlates with *linguistic structures* at or above the word level. How can we use it for *word recognition*?
- Approach: model prosody associated with *hidden events,* such as sentence boundaries and disfluencies, and include this information in the language model [1].

[1] A. Stolcke, E. Shriberg, D. Hakkani-Tur, and G. Tur, "Modeling the Prosody of Hidden Events for Improved Word Recognition", Proc. EUROSPEECH 1999, vol. 1,pp.307-310, Budapest.

SRI International

Modeling Approach

- F: prosodic features associated with acoustics A
- E: event sequence behind word sequence W

$$
\begin{aligned}
W_{\text{best}} &= \text{argmax}_W\, P(W \mid A, F) \\
&= \text{argmax}_W\, P(W \mid F)\ \mathrm{P}(A \mid W, F)\, /\, \mathrm{P}(A \mid F) \\
&\approx \text{argmax}_W\, P(W, F)\ \mathrm{P}(A \mid W) \\
&= \text{argmax}_W \textstyle\sum_E P(W, E, F)\ \mathrm{P}(A \mid W)
\end{aligned}
$$

Example of different event sequences for the same W :

Right **<S>** *I* **<REP>** *I don't* **<DEL>** *uh* **<FP>** *I'm not sure* **<S>**

Right, I **<REP>** *I don't* **<S>** *uh* **<FP>** *I'm not sure* **<S>**

SRI International

Hidden Event Modeling

$$P(W, E, F) = P(W, E)\, P(F \mid W, E)$$

$$\approx P(W, E) \prod_{i=1..n} P(F_i \mid W, E_i)$$

- $P(W, E)$ modeled using a standard N-gram
- During testing event sequence E is unknown. Need to sum over all possible event sequences.
- Joint model $P(W, E, F)$ equivalent to HMM with:
  - (word,event) pairs as states
  - transition probabilites provided by the N-gram
  - emission probabilities provided by prosodic model

Prosodic Model

$$P(F_i \mid W, E_i) = P(F_i \mid W)\, P(E_i \mid F_i, W) \,/\, P(E_i \mid W)$$

$$\approx P(F_i)\, P(E_i \mid F_i, W) \,/\, P(E_i)$$

- Prosodic features are independent of the word identity. We only make use of the alignment information associated with W to extract the prosodic features.
- We train CART decision trees to estimate the posterior probabilities $P(E_i \mid F_i, W)$.
- $P(F_i)$ treated as a constant.
- $P(E_i)$ constant by equating priors on training set.

SRI International

Experiment

- Used 5 hidden event types: sentence boundary, filled pause, repetition, deletion, no-event. Only 18% of word boundaries have non-null events!
- LDC event-annotated conversations were used to obtain a LM; used to automatically annotate the rest of the SWBD training corpus. The whole event-labelled corpus was then used to train a 4-gram hidden event LM, interpolated w/BN.
- *Decision tree* trained using the devel2001 data.
- *Prosodic features*: previous/current pause durations, turn info, last rhyme/vowel durations in word, F0 patterns.
  - Tree accuracy: 61.8%
  - Tree efficiency: 42.5%
- Optimized prosodic model and hidden event-model weights.

Results

Baseline 1: PLP MMIE CW adapted acoustic models, pronunciation probabilities, 4-gram LM (without classes)

Baseline 2: added class 4-gram, duration & pause models

	dev2001	eval2001	eval2002
Baseline1	27.99	26.61	29.00
+ HE LM	27.92	26.47	28.82
+ prosody	27.72	26.35	28.67
Baseline 2	27.19	26.05	28.15
+ HE LM	27.08	26.03	28.16
+ prosody	26.90	25.86	27.99

SRI International

Conclusions

- Prosody can be leveraged for LVCSR via modeling of linguistic structures
- Case in point: inter-word “hidden events”.
- 0.3% abs. improvement over standard knowledge sources.
- Small additional improvement with other prosodic models (phone durations, pause LM).