Case study: ASR of multiparty conversations

Steve Renals

Automatic Speech Recognition— ASR Lecture 14
19 March 2009

Overview

Transcription of speech in meetings

- Large vocabulary continuous speech recognition
- Speaker independent, conversational style, environment with reverberation, multiple acoustic sources: An "ASR complete" problem
- Applications: transcription, summarization, translation, ... of meetings, lectures, seminars, ...
- Development of a system
  - language resources
  - baseline system
  - acoustic models
  - language models and vocabulary

Example

Right I didn't mean to imply that
Yeah
that we - that we shouldn't discuss this now, but I'm - I'm just saying that
Oh not right now, but I mean in the future. So at this meeting with Liz
Right
I - you know - I mean
Right
I - I do - I'd like to - I like that stuff
Sure sure
So when is she showing up?
Well, I mean, they're coming in April
April. OK
Right. But, um

>

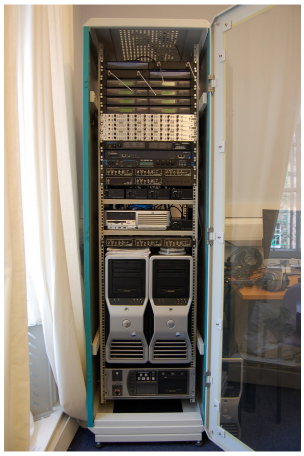
Spontaneous conversational speech

Substantial segmental and suprasegmental variations not found in read speech:

- Variations in intonation (F0) and timing (segment durations)
- Hestitations
- False starts
- Ungrammatical constructs
- Increased expression (eg laughter)

Instrumented meeting rooms

- Capture all aspects of a "communication scene"
- Four a four-person meeting instrument a room with:
    - 6 cameras (4 close-up, 2 room view)
    - headset microphones
    - microphone array (distant microphones)
    - capture of data projector, whiteboard, handwriting (digital pens)



Baseline system: conversational telephone speech (CTS)

- General strategy: build a baseline system using CTS data, than adapt to meetings data
- GMM/HMM system: cross-word, state-clustered triphone models (7 000 states, 16 Gaussians/state)
- PLP front end: MF-PLPs + zeroth cepstral coefficient + first derivatives + second derivatives
- Cepstral mean and variance normalization:
    - over a complete recording normalize the cepstral coefficients by subtracting the mean vectors and dividing by the variance
    - reduces distortion by removing channel effects (spectral characteristics of microphone)
    - channel effects are multiplicative in spectral domain, additive in cepstral domain
- VTLN
- Constrained MLLR speaker adaptive training

CTS: Pronunciation dictionary

- Pronunciation dictionary based on UNISYN (115 000 words)
- Added a further 11 500 domain specific pronunciations
- Automatic pronunciation generation (using Festival), then hand corrected
- Accuracy of automatically generated pronunciations
  - In vocabulary: 98% phone accuracy, 89% word accuracy
  - New words: 89% phone accuracy, 51% word accuracy
- Final vocabulary of 50 000 words derived from training data and language model sources
- Language models constructed from about 1 200 million words: transcripts, web-retrieved texts based on in-domain n-grams, broadcast news transcripts, newswire

CTS: Accuracy

- Using the the NIST 2001 evaluation data
- Pass 1 (no VTLN, no MLLR): 37.2% WER
- Pass 2 (VTLN, no MLLR): 33.8% WER
- Pass 3 (VTLN, MLLR): 32.1% WER

Meeting corpora

- Existing, well-studied speech corpora: conversational telephone speech (CTS), broadcast news (BN) — hundreds/thousands hours of transcribed speech data
- Transcribed meeting collections
  - ICSI meeting corpus (70 hours): 3–12 person meetings, audio only
  - AMI meeting corpus (100 hours): mainly 4 person meetings, multimodal
  - Some other smaller corpora
- Meeting transcription tasks
  - Conference room or lecture
  - Headset microphones or microphone array

Statistics of meeting corpora

Meeting resource	Avg Dur (sec)	Avg. Words/Seg
ICSI	2.11	7.30
NIST	2.26	7.17
ISL	2.36	8.77
AMI	3.29	10.09
VT	2.49	8.27
CHIL	1.80	5.63

- Segment is speech with no silence of more than 100ms
- Average utterance durations greater than CTS, more variation in duration

Meeting corpus OOV rates (%)

	Meeting resource specific				"Padded"			
	Vocabulary Source				Vocabulary Source			
Corpus	ICSI	NIST	ISL	AMI	ICSI	NIST	ISL	AMI
ICSI	0.00	4.95	7.11	6.83	0.01	0.47	0.58	0.57
NIST	4.50	0.00	6.50	6.88	0.43	0.09	0.59	0.66
ISL	5.12	5.92	0.00	6.68	0.41	0.37	0.03	0.57
AMI	4.47	4.39	5.41	0.00	0.53	0.53	0.58	0.30
ALL	1.60	4.35	6.15	5.98	0.16	0.42	0.53	0.55

- Meeting resource specific: vocabularies derived from training data
- Padded: vocabularies extended to 50 000 words using most frequent additional words from broadcast news

Audio preprocessing

- Segment audio, discarding silence and noise
- Label speakers for adaptation
- Normalize input channels
- Suppress noise and cross-talk
- For headset microphones main problem is the elimination of cross-talk:
  - use specific features for cross-talk suppression: cross-correlation, cross-channel energy, signal kurtosis
  - train a classifier to detect speaker activity: MLP with 101 frames (1s) of input context

Language models

- n-gram language models (4-grams)
- Small amount of in-domain text data
- Augment this with:
  - other conversational speech transcripts
  - broadcast news
  - data retrieved from the web using n-grams from meeting data as queries
- Perplexities on meeting data
  - Trigram: 84
  - 4-gram: 81

Language model data sources

LM component	size	weights (trigram)
AMI data (prelim.)	206K	0.038
Fisher	21M	0.237
Hub4 LM96	151M	0.044
ICSI meeting corpus	0.9M	0.080
ISL meeting corpus	119K	0.091
NIST meeting corpus	157K	0.065
Switchboard/Callhome	3.4M	0.070
webdata (meetings)	128M	0.163
webdata (fisher)	128M	0.103
webdata (AMI)	138M	0.108

Adaptation from CTS to ICSI Meetings

- Testing on development test data from ICSI corpus...
- CTS system (255 hours training data): 33% WER
- Trained on 70 hours ICSI data (in domain): 25.3% WER
- MAP adapted CTS models: 24.6% WER
- Technical issue: CTS is narrowband data, meetings are wideband. One iteration of MLLR transforms were used to estimated the narrowband/wideband transform (using ICSI data)

Example (ASR)

right yeah race i didn't mean imply that that we'd did that we should that that's just now but i'm i'm saying that
oh not right now i mean in the future
right
so at this meeting with with you know i mean
right
i i do i'd like to i'd like to stop
sure sure
when she showing
well i mean theyre coming in april
april but in right
right but

>

Summary

- Putting the pieces together to build a large vocabulary system for conversational speech
- Adapting to a new (but related) domain
- Accuracies on test data
- Next lecture: robust speech recognition
- References: Renals, Hain and Bourlard (2007); Hain et al (2005)
- Development of Broadcast News transcription system: Woodland (2002)
- A recent CTS system: Chen et al (2006)