The ISL evaluation system for RT-03 CTS

Hagen Soltau, Hua Yu, Florian Metze,
Christian Fügen, Qin Jin, Szu-Chen Jou

Interactive Systems Laboratories
Universität Karlsruhe, Carnegie Mellon University

Overview

- Acoustic Modelling
  - VTLN, FSA-SAT, STC
  - Clustering across phones
  - MMIE via confusion networks
- Language Modelling
- Decoding Strategy

Acoustic data

- AM training
  - 265h SWB+CHE (with ISIP transcripts)
  - 32h CELL
  - 65h CTRAN
- Development sets
  - dev01 : 1h subset from eval01 (manual segmentaion)
  - dry-run : 1h subset from eval02 (automatic segmentation)
  - Error rates on dev01 unless stated otherwise

Corpus Preparation

- Frames with null energy
  - yield extreme likelihoods for SIL model
  - hurt FSA-SAT reestimation
  - removing those frames : 33.4% -> 32.8%
- Removed all one-word segments
  - Contain a lot of noise / silence
  - Unprecise segment boundaries
  - 37.1% -> 36.4%

Corpus Preparation II

- Resegmentation
  - Limited silence at begin and end of turns : 15 frames
  - Discarded ~11h of silence
  - 35.1% -> 34.8%
- Likelihood checking
  - Discarded turns with poor likelihoods: ~20h
  - 32.4% -> 32.2%

Front-end

- Basic features:
  - 13 cepstra (including C0)
  - Context +/- 5 frames
  - LDA : dimension reduction 143 -> 42
  - CMS / CVN per conversation side
- MFCC vs. PLP (setup without VTLN & MLLR)
  - MFCC : 39.0%
  - PLP : 40.7%

VTLN

- Old approach:
  - Extract CMS/CVN with initial warp factor
  - Compute likelihood for each warp factor with FIXED CMS/CVN
  - Mismatch between channel normalization and frequency warping
  - VTLN gain:
    - Without MLLR: 39.0 -> 36.1
    - With MLLR : 33.5 -> 31.8

VTLN II

- Revised VTLN strategy:
  - Take the correct CMS/CVN for each warp factor
  - Brent search
  - Interleaved estimation of VTLN, FSA, MLLR
  - Results:
    - Old VTLN : 33.2
    - New VTLN : 32.4

Clustering

- Entropy based divisive clustering
- Wide contexts :
  - Quintphones : 34.7
  - Septaphones : 34.2
- Stress tags :
  - Dictionary used stress tags for vowels
  - Inconsistencies due to dictionary expansion
  - Removing tags : 32.7 -> 31.8

Clustering II

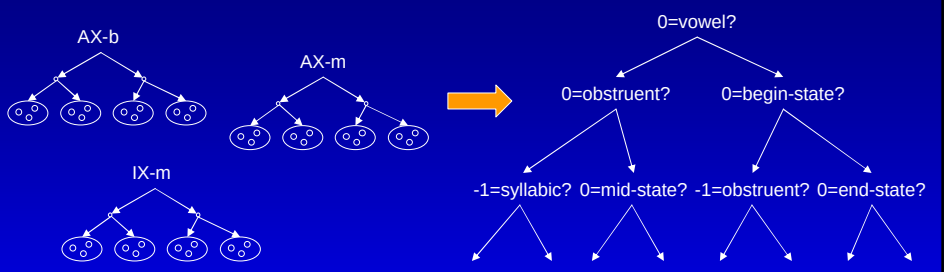
- Semi-tied models
  - Tree 1 : means and covariances (10k leafs)
  - Tree 2 : mixture weights (50k leafs)
  - Results:
    - 10k codebooks + 10k distributions : 32.8
    - 10k codebooks + 50k distributions : 31.8

    (5% increase of model parameters)

Alternative Clustering [HuaYu 2003]

- Standard way :
  - Grow tree for each context independent HMM state
  - 50 phones, 3 states : 150 trees
- Alternative : clustering across phones
  - Global tree -> parameter sharing across phones
  - Computationally expensive to cluster -> 6 trees
    (begin, middle, end for vowels and consonants)
  - Quintphone context

Clustering across phones II

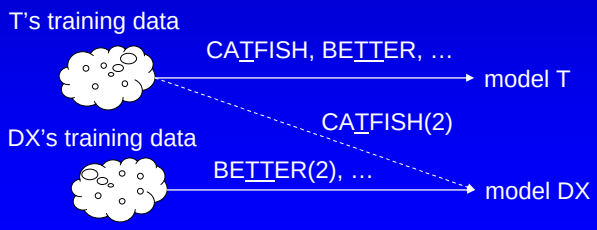


- Allows better parameter tying (tying now possible across phones and sub-states)
- Alleviates problems in a lexicon: over-specification and inconsistencies → no need for an optimal phone set, preferable for multi-lingual / non-native speech recognition
- Implicitly models subtle reduction in sloppy speech

Side Effects of Pron. Variants

- Increases lexical confusability for decoding
- Contaminates model during training, if a variant is spurious

CATFISH K AE T F IH SH
CATFISH(2) K AE DX F IH SH



Clustering across phones : Experiments

- Cross-substate clustering doesn't make any difference
- Cross-phone clustering w/ 6 trees: vowel-b, vowel-m, vowel-e, consonant-b, consonant-m, consonant-e
- Single pronunciation lexicon has 1.1 variants per word (instead of 2.2 variants per word)

Dictionary	Clustering	WER 66hr training set	WER 180hr training set
multi-pronunciation	traditional	34.4	33.4
	cross-phone	33.9	-
single pronunciation	traditional	34.1	-
	cross-phone	33.1	31.6

Results are based on first pass decoding on dev01

Semi-tied full Covariances

- On top of LDA
- Phone dependent STC classes
  - Work for unadapted models : 38.5 -> 34.4
  - Do not work in combination with MLLR even if identical MLLR regression classes and STC classes
  - Global STC for adapted models: 34.1 -> 32.2
- Re-estimation of STC for each test speaker
  - 26.8 -> 26.6

FSA-SAT

- On top of LDA / STC
- Global FSA per conversation side
- Null energy frames exluded for FSA estimation
- Gain: 28.9 -> 27.8

MMIE training

- Accumulate statistics:
  - Convert lattices to confusion networks
  - Run full forward/backward over networks
  - Same performance as collecting statistics via fwd/bwd for each node of the lattice
- Update as in [Woodland&Povey2000]
- Results:
  - Small setup (30h training) : 41.9 -> 40.9
  - Full setup (STC,FSA-SAT,MLLR) : 28.3 -> 27.6

Training Procedure

1. Train fully-continuous models (10k codebooks)
   - Re-organize data per HMM state (fixed alignment)
   - Speeds up the training drastically
   - Grow mixture components (~ 30 iterations)
   - Estimate STC (4 iterations)
2. Expand to semi-continuous models (50k distribs)
   - FSA-SAT viterbi training (4 iterations)
   - MMIE training (1 iteration)

Vocabulary

- Vocabulary
  - 41k vocabulary selected from SWB, BN, CNN
  - CNN used for vocabulary selection but not for LM training
- Pronunciation Variants
  - Rule derived dictionary expansion : 95k entries
- Probabilities
  - Based on frequencies (forced alignment)
  - Viterbi decoding : probabilities as penalties (e.g. max = 1)
  - Confusion networks : real probabilities (e.g. sum = 1)

Language Modelling

- Better text processing, more data:
  - Removing inconsistencies : 32.6 -> 32.4
  - Adding CELL + CTRAN transcripts : 32.4 -> 31.6
- LM interpolation (context dependent)
  - 3gram SWB : 31.4
  - 3gram SWB + 5gram class SWB : 31.0
  - 3gram SWB + 5gram class SWB + 4gram BN : 30.3
  - 3gram SWB + 5gram class SWB + 4gram BN + 4gram CNN : 30.5

Segmentation

- Approach (details tomorrow -> Qin Jin)
  - Initial energy based segmentation
  - Train GMM for speech / silence
- Oversegmentation for ASR
  - Segmentation used for ASR is not identical with the one submitted for MDE
  - F10 : 14.3% segmentation error, 41.9% WER
  - F11 : 9.7% segmentation error, 43.0% WER

Decoding Engine

- Single pass decoder
- Minimized search network
  - build initial tree structured graph
  - compress network via merging of isomorph subgraphs
- Full language model lookahead
- linguistic states via polymorphism
  - exploits subgraph dominance automatically

Decoding strategy

(results on dry-run, automatic segmentation)

pass 0 : 35.0 : Tree-150, MMIE, STC-50, 3gram SWB
pass 1 : 30.7 : Tree-150, ML, STC-1, VTLN, lattice-MLLR, smallLM
pass 2 : 28.5 : Tree-150, ML, STC-1, VTLN, hypo-MLLR, bigLM
pass 3 : 27.7 : Tree-150, ML, STC-1, FSA-SAT, bigLM
pass 4 : 27.2 : Tree-150, MMIE, STC-1, FSA-SAT, bigLM
pass 5 : 26.6 : Tree-6, ML, STC-1, FSA-SAT, bigLM, SPDict
pass 6 : 26.2 : Tree-150, MMIE, STC-1, FSA-SAT, bigLM
pass 7 : 26.4 : re-adapted Tree-6 models, 8ms frameshift
pass 8 : 25.4 : re-adapted Tree-150 models, 8ms frameshift
pass 9 : 24.9 : system combination

Decoding Strategy II

- System Combination
  - Combine tree-150, tree-6, 8ms, 10ms output
  - Confusion networks over multiple lattices and rover
  - Confidences computed from combined confusion networks
  - Best single output (Tree-150) : 25.4
  - CNC+rover : 24.9
- Results on eval03
  - Tree-150 single system : 24.2
  - CNC+rover : 23.4

Decoding Time

- Full system : 190 rtf on 2.4Ghz, 550MB
- single adapted pass, reducing beams

Pruning parameter	RTF (on P4 2.4Ghz)	WER (on eval 03)
Beam = 2.2 (eval mode)	12.0	24.2
Beam = 1.5	4.7	24.6
Beam = 1.3	2.9	25.1
Beam = 1.1	1.4	26.0
+ transN = 35	1.0	26.1
+ delayed interpolation	0.9	26.1

Summary

- last participated in 1997
  - revived 97 system in January 2002 : 35.1%
  - Summer 2002 : 29.2%
  - April 2003 : 21.8%
- 23.4% WER on eval-03 (24.7% on dry-run)
- Stable setup for VTLN, FSA, STC, MLLR
- Cross-adaptation of Tree-150 and Tree-6