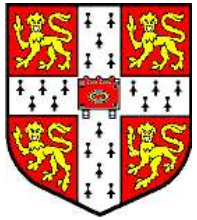
Experiments with Fisher Data

Gunnar Evermann, Bin Jia, Kai Yu, David Mrva
Ricky Chan, Mark Gales, Phil Woodland

May 16th 2004



Cambridge University Engineering Department

Overview

- Introduction
- Pre-processing 2000h of Fisher data
- Fisher dev04 test set
- Language Modelling
- Acoustic model training on Fisher
- Modelling techniques (MMI prior for MPE, MPE-MAP, Gaussianisation)
- Conclusions

Fisher Data Processing

- Original transcriptions: 1940h data (1758h BBN data, 182h LDC data)
- Normalise the text, join segments, pad with silence as necessary
- Apply replacement rules
  - Abbreviations, typos, non-speech, etc.
  - e.g. CD → C. D., PRIVELAGE → PRIVILEGE, [STATIC] → -
  - about 11k replacement rules were produced
- Produce pronunciations for 6800 unknown words (4100 whole words and 2700 partial words) with frequency greater than 2
- 8500 unknown words remain → remove 14h worth of segments.
- Align the segments and normalise silence boundaries
  - <30h segments failed to align
  - 1819h data remained (Gender imbalance: 1042h female, 777h male)

Training and Test Sets

- Acoustic training data

  h5train03b 360h data set
  - 290h LDC data (Swb1, CHE, Swb Cellular) with MSU/LDC careful transcriptions.
  - 70h BBN data (Cellular, Swb2-2) with quick transcriptions

  fisher3896 520h Fisher data set, 3896 conversations with "Algorithm 1" quick transcriptions: results presented in St.Thomas

  fsh2004 1820h Fisher data set

  fsh2004sub 400h Fisher subset (balanced for gender and line condition)

  fsh2004sub2 800h Fisher subset (gender balanced)

- Test sets

  eval03 6h set from Fisher and Swb2-5 data, 72 conversations

  dev04 3h set from Fisher, 36 conversations

CTS dev04 test set

- Ran CU-HTK 2003 10xRT system on dev04 to test robustness on Fisher

pass	eval03Fi	dev04
P1	29.7	29.8
P2 latgen	20.0	21.0
P3 (SAT)	18.8	19.3
P3 (SPron)	18.9	19.5
final	18.4	18.9
final (STM)		18.6

%WER on two Fisher test sets (3h each) with 2003 10xRT system

- LM perplexity with RT03 fourgram: eval03Fi: 65.7 dev04: 61.9
- Overall dev04 is slightly harder than eval03Fisher and the progress set (18.2%)
- Would like to know gender and line types for dev04 sides

How (not) to Optimise LM Interpolation Weights

- Train separate n-gram on each corpus (Swb1, Cell1, Fisher, BN, Google, etc.)
- Optimise interpolation weights on a dev set (reference STM)
- Merge component n-grams into single LM
- Problem: reference STM had all contractions expanded (`don't`→`do not`)

corpus	size	weight (STM)	weight (non-exp)
BN	427M	0.137	0.120
google	63M	0.071	0.063
cell1	0.2M	0.230	0.021
che/sw1	3M	0.022	0.042
swb2	0.9M	0.006	0.053
fisher	21M	0.534	0.700

weights optimised on dev04

Fisher Language Models – Perplexities

- Train separate word 4-gram on all fisher data (21M words)
- Interpolate with RT-03 component n-grams

Language Model	optimised on	Perplexity
fgint03	dev01+eval01/03 exp	62.0
fgint04	dev04 exp	53.6
fgint04	dev04 no exp.	52.8

Perplexities of word 4-grams on **dev04** with unexpanded contractions

fgint03 word fourgram used in 2003 CU-HTK system (5 components)

fgint04 above components plus fsh2004 4-gram component

- size of fgint04: 6.3M bigrams, 11.6M trigrams, 4.8M 4-grams

Fisher Language Models – WER

- Tested new LM by rescoring 2003 CU-HTK full system lattices

LM	optimised on	WER	Swb	Fsh
fgint03	dev01+eval01/02 exp	23.5	27.4	19.3
fgint04	dev04 exp	22.6	26.7	18.3
fgint04	dev04 noexp	22.6	26.8	18.1
fgint04	dev04+eval03 noexp	22.6	26.8	18.1

%WER on eval03, rescoring 2003 CU-HTK system lattices
(fgintcat03, adapted HLDA MPE models)

On the Fisher portion of the test set:

- Using Fisher data for language modelling gives 1.2% abs.
- Optimising interpolation weights incorrectly cost 0.2% abs.

Fisher acoustic modelling

Overall strategy:

- Pre-process **all** data (align, VTLN, etc.)
- Fix various issues with Software & infrastructure for large data sets (issues with numerical accuracy, avoid having directories with 20k files, etc.)
- Select manageable subset as baseline for investigation of new techniques 400h, balanced for gender, line conditions, topics
- Concurrently investigate training on larger amounts of data
  - MLE & MPE models for 800h fisher set
  - MLE models for all fsh2004 + h5train03b (2200h total)

Subset selection

A 400h subset was selected from the whole fisher data set

- only whole conversations used
- only use sides for which all labels (gender, line, topic) were available
- ignore sides that were too short or had a high percentage of data fail to align
- balance gender
- select 25% cellular data (like current and progress sets)
- aim for even topic distribution

MLE/MPE on 400h Fisher

- Train models on new 400h Fisher subset
- Number of parameters same as before (about 6000 states, 28 components)

			eval03	eval03Sw	eval03Fi	dev04
ML	h5train03b	(360h)	31.7	36.1	27.1	28.1
ML	fisher3896	(520h)	30.8	34.7	26.6	26.9
ML	fsh2004sub	(400h)	30.8	34.6	26.7	26.8
MPE	h5train03b	(360h)	27.3	31.6	22.7	23.7
MPE	fisher3896	(520h)	26.2	30.0	22.2	22.3
MPE	fsh2004sub	(400h)	26.3	29.9	22.5	22.3

%WER on eval03 and dev04, unadapted, 2003 trigram

- New Fisher 400h set gives very similar performance to old 520h one
- WER reduction of 1% abs. over 2003 training set

MPE with dynamic MMI prior

- Use dynamic MMI estimates instead of ML estimates as the I-smoothing prior
- 4 sets of statistics to accumulate: num, den, ml, mmi-den, extra 1/3 memory and disk space, no extra computation

MPE Prior	MPE-τ^I	MMI-τ^I	eval03	eval03Sw	eval03Fi
Dynamic ML	50	—	26.3	29.9	22.5
Dynamic MMI	75	0	25.9	29.6	21.9

%WER on eval03 for MPE models trained on fsh2004sub, unadapted, 2003 trigram

Larger data sets

- Compare 400h subset with larger training sets

			eval03	eval03Sw	eval03Fi	dev04
ML	h5train03b	(360h)	31.7	36.1	27.1	28.1
ML	fsh2004sub	(400h)	30.8	34.6	26.7	26.8
ML	fsh2004sub2	(800h)	30.5	34.4	26.4	26.5
ML	fsh2004h5train03b	(2200h)	30.2	34.1	26.0	26.4
MPE	h5train03b	(360h)	27.3	31.6	22.7	23.7
MPE	fsh2004sub	(400h)	25.9	29.6	21.9	21.9
MPE	fsh2004sub2	(800h)	25.1	28.9	21.1	21.3

%WER on eval03 and dev04, unadapted, 2003 trigram, Fisher models used MMI prior

- Adding 1800h of Fisher to acoustic training improves ML models by 1.5% abs.
- 2.2% abs. WER reduction from using 800h Fisher instead of 360h h5train03

Putting it all together: CU-HTK P1-P2 System (5xRT)

			eval03	eval03Sw	eval03Fi
h5train03b	(360h)	LM03	24.6	28.7	20.2
h5train03b	(360h)	LM03 + fsh	23.3	27.6	18.6
fsh2004sub	(400h)	LM03 + fsh	22.7	26.7	18.4
fsh2004sub2	(800h)	LM03 + fsh	22.0	25.9	17.8

%WER on eval03, MPE models, word 4-gram, simple adaptation

- h5train03b: Fisher data in LM gives 1.3% abs. improvement (1.6% on Fisher)
- fsh2004sub (400h) performs 0.6% better than h5train03b (360h)
- doubling the amount of fisher data gives an additional 0.7%
- Total WER reduction of 2.6% abs. (2.4% on Fisher) from using 800h of Fisher data instead of 360h Swb/CHE data for acoustics and LM

MPE Training for Gender-dependent Models

- GD MPE training of means and mix weights on top of GI MPE training
- Static MPE-GI model parameters used as the I-smoothing prior

Unadapted single pass decode:

System	MPE Prior	eval03	Male	Female
MPE-GI	Dynamic MMI	25.9	27.3	24.5
MPE-GD	MPE-GI model	25.6	27.1	24.1

%WER on eval03, fsh2004sub models, unadapted, 2003 trigram

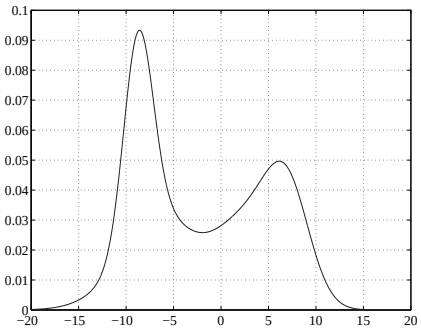
Test with adaptation in P1-P2 system:

System	MPE Prior	eval03	Male	Female
MPE-GI	Dynamic MMI	22.7	24.0	21.4
MPE-GD	MPE-GI model	22.4	23.8	21.0

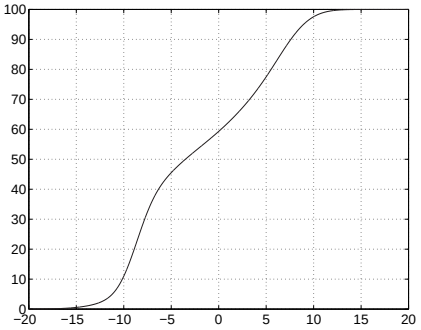
%WER on eval03, fsh2004sub models, adapted, LM03+Fsh 4-gram, P1-P2 system

Gaussianisation

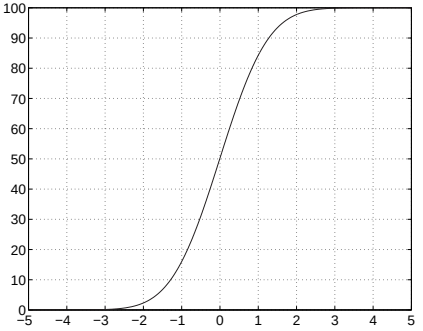
- Transform any distribution to standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$



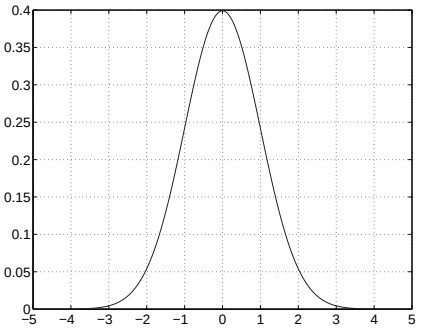
Source PDF



Source CDF



Target CDF



Target PDF

- Use multiple-stream (one per dimension) GMMs per speaker after HLDA:
  - simplified version of iterative Chen and Gopinath scheme;
  - more compact, smoother, representation than using data (IBM style);
  - simple to implement in HTK ...

- May be viewed as higher-moment version of CMN and CVN

Gaussianisation – MPE Results

- `fsh2004sub` (400hr) training set - 28 components + `varmix`;
- unadapted decode with 2003 trigram

System	Swb	Fsh	Tot
Baseline	29.7	21.9	26.0
+CN	28.8	21.3	25.2
Gaussianised	29.8	21.9	26.0
+CN	28.7	21.3	25.1
CNC	28.1	20.8	24.6

- No gain over baseline with `fsh2004sub`
  disappointing - with h5train03b 0.4% absolute gain on `eval03`
- Possibly useful for system combination (but need adapted numbers)

Conclusions

- Fisher data for LM training reduces WER by 1.3% abs.
- Overall 2.6% WER reduction in P1-P2 system from using 800h Fisher for acoustic training and all Fisher data for LM
- Using all Fisher and h5train03b together in MPE should improve WER further (0.3% in ML)
- Need to investigate number of model parameters for large training sets