NIST 2006
Speaker Recognition Evaluation
QUT Submission

Speaker: Brendan Baker
bj.baker@qut.edu.au

Speech and Audio Research Laboratory
Queensland University of Technology

QUT TEAM: Robbie Vogt, Brendan Baker

System Overview

- Attempted the [1, 3, 8]-conv4w training, 1-conv4w testing conditions. In results tables, submission is designated as **QNI. (QUT 'N IBM).**
- The final system was a combination of 8 sub-systems
  - Cepstral GMM (including Session Variability modelling)
  - Text-constrained (Syllable) Cepstral HMM (1conv4w only)
  - Idiolect N-gram
  - Phonetic N-gram
  - Prosodic Gesture N-gram
  - **Session Variability GMM Supervector SVM** (1conv4w only)
  - **Phonetic Binary Decision Tree**
  - **GMM index Binary Decision Tree**

  The systems indicated in **red** were contributed by IBM
- Multiple score outputs provided by some of these systems. See system description document for details.

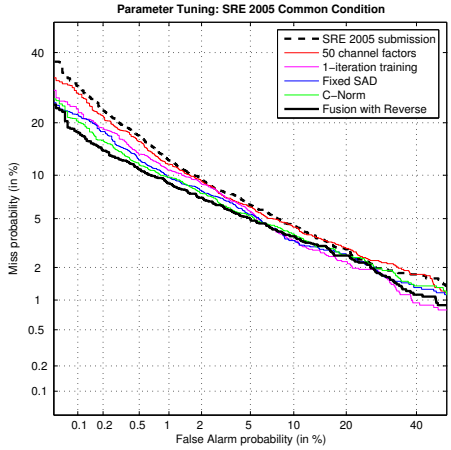
Development Data

- Background Data was drawn from SRE 2004 and Switchboard-II for most systems.
- Normalisation data was also drawn from SRE 2004, the same normalisation scripts were used for all systems.
- System parameter tuning and fusion training were performed on the SRE 2005 data and protocol.

Cepstral GMM System

- Very similar to the cepstral GMM system submitted last year, with some continued parameter tuning and tweaking.
  - GMM-UBM verification structure (512 mixtures)
  - Session variability modelling
  - Feature warped MFCCs with deltas
  - Evaluated the protocol in both forward and reverse directions for the core condition
    - Reverse → train on test segments; test on train segments.
    - Forward only for 3- and 8-side training
  - CT-Norm and ZT-Norm normalisation
    - C-Norm for forward, Z-Norm for reverse
- See the system description for details

System Parameter Tuning



- Evolutionary development from SRE'05 system
  - 20 → 50 session factors
  - 5 → 1 iteration of speaker model training
  - Fixed a bug in our SAD, retrained subspace
  - Z-Norm → C-Norm
  - Adding reverse testing
- Min DCF: 0.215 → 0.155
  - 28% rel. improvement

Single Iteration Training

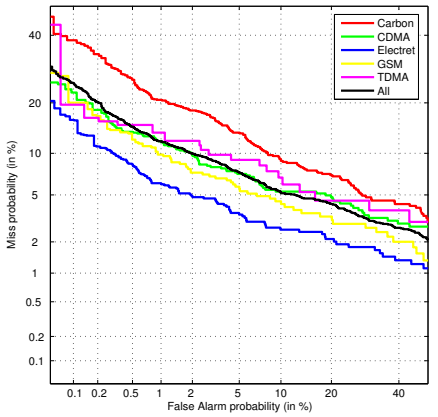
- Previously QUT has advocated **multiple-iteration** MAP adaptation for speaker model training †
  - MAP adaptation within a E-M framework
  - Theoretically more accurate models and practically more optimal w.r.t. the MAP criterion
  - Also provided better performance
- This trend is reversed for session variability modelling where a single iteration works better ‡
  - Better matches the **testing** procedure, where a single-pass adaptation is used to estimate the session factors
    - Multiple iterations are still better for training the session subspace

† Pelecanos, Vogt, Sridharan, "A study on standard and iterative MAP adaptation for speaker recognition," SST, 2002

‡ Vogt, Sridharan, "Experiments in session variability modelling for speaker verification," ICASSP, 2006

Effect of Test Channel Type

Performance by Test Segment Channel Type: SRE 2005



Test Type	Min DCF	EER
Electret	0.0128	4.03
GSM	0.0168	5.35
CDMA	0.0176	6.51
TDMA	0.0178	8.13
Carbon	0.0289	9.59
All	**0.0195**	**6.5**

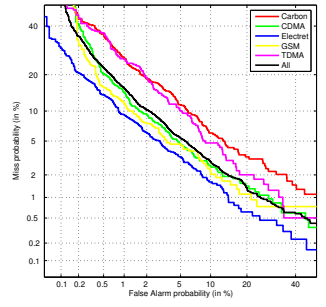
- Still a considerable difference between Carb and Elec even with session variability modelling!
  - Cellular types quite consistent.
  - Channel labels **courtesy of IBM's channel detector**

Effect of Test Channel Type

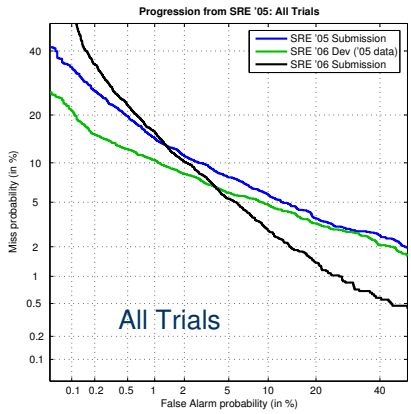
- Investigated the use of C-Norm instead of Z-Norm to compensate for this difference
  - This gave a small (5% rel.) improvement to the overall performance.
  - The large discrepancy between the different channel types persists.
  - Similar for the '06 results.

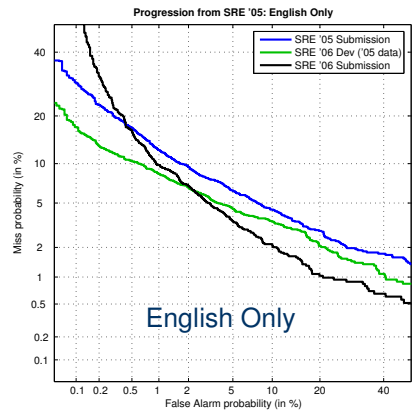
Test Type	Min DCF	EER
Electret	0.0181	3.86
GSM	0.0204	4.63
CDMA	0.0235	4.88
TDMA	0.0345	7.29
Carbon	0.0347	7.93
All	**0.0254**	**5.23**

Performance by Test Segment Channel Type: SRE 2006



Comparison of ’05 and ’06 Systems





Text-constrained Cepstral HMM

- Similar to system used in QUT2005 submission†
- Only used for the 1conv4w submission.
- Multilingual text-constrained framework used for segmentation & modelling.
  - Pseudo-syllabic segmentation process (broad phone recogniser used)
  - Modelling and recognition constrained to syllabic events
  - Allows for substitution of features & modelling paradigms

† Baker, Sridharan “Speaker Verification using Hidden Markov Models in a Multilingual Text-constrained Framework,” Speaker Odyssey 2006

Changes from last year…

- A number of changes were made in the hope of improving performance and giving more stable results.
- Front-end phone recogniser:
  - 4 broad classes → 6 broad classes (more vowel resolution)
  - OGI training → Callhome training
- Scoring
  - No score norm → T-Norm
- Score combination
  - Fuse all 216 into 1 → Tiered fusion (216 into 6 into 1)
  - SVM → LLR

Text-constrained Performance

- Fairly consistent results for EN and ALL trials condition.
- Better EER for SRE06(EN) than DEV(EN)

Test Type	Min DCF	EER
1conv4w		
DEV (EN)	0.0476	9.57
DEV (ALL)	0.0484	9.99
SRE06 (EN)	0.0483	9.27
SRE06 (ALL)	0.0529	10.66

Other finds:

- T-Norm was found to help a little! C-Norm or Z-Norm probably needed!
- Increase to 6 phone classes didn't really help. Inappropriate class definitions?
- A lot more analysis/development needs to be done on this system!

Lexical System

- Based on Doddington's word n-gram speaker recognition work and very similar to QUT 2005 lexical system. †
- Additions/Changes from 2005:
  - Score Normalisation: Z & T-Norm used
  - Modelling: Used Bag-of-bigrams and Bi-grams
- See system description for modelling/scoring details

† Baker, Vogt, Mason, Sridharan "Improved Phonetic and Lexical Speaker Recognition through MAP Adaptation", Speaker Odyssey 2004

QUT Queensland University of Technology Brisbane Australia
a university for the real world
Speech and Audio Research Laboratory

Lexical System Performance

- Development results showed improvement from SRE05 (tuning & norm).
- Degradation in performance in SRE06 compared to DEV06.
  Why?
  - Increase in non-english data? Or due to different BBN ASR?
  - Bag of 2grams outperforms Bigrams
- Slight gains from fusion of modelling types.

	BAG OF 2 GRAMS		BIGRAMS	
Test Type	**EER**	**MinDCF**	**EER**	**MinDCF**
1conv4w (EN)				
SRE 05	27.69	0.0958	-	-
DEV 06 (05 Data)	26.89	0.0915	32.75	0.0963
SRE 06	27.29	0.0915	32.44	0.0955
1conv4w (ALL)				
SRE 05	28.55	0.0947	-	-
DEV 06 (05 Data)	27.64	0.0906	33.67	0.0967
SRE 06	31.15	0.0951	35.99	0.0978
8conv4w (ALL)				
SRE 05	14.52	0.0600	-	-
DEV 06 (05 Data)	13.57	0.0530	17.72	0.0613
SRE 06	16.12	0.0702	19.85	0.0737

Phonetic N-gram Systems

Extended QUT 2005 Phonetic N-gram System. †

2006 Updates…

- 2 Different PPRLM systems tested this year.
  - OGI trained and Callhome trained
- 2 modelling variations used
  - Bigrams & Bag-of-trigrams
- Score stream combination
  - Logistic regression used to calculate optimal stream weights.†† More stable than SVM?
- Score normalisation added
  - T-Norm and Z-Norm

† Baker, Vogt, Mason, Sridharan "Improved Phonetic and Lexical Speaker Recognition through MAP Adaptation", Speaker Odyssey 2004

† † Niko Brümmer, "FoCal: Tools for Fusion and Calibration of automatic speaker detection systems," available at http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm, 2005

QUT Queensland University of Technology Brisbane Australia
a university for the real world
Speech and Audio Research Laboratory

Phonetic System Performance

- Callhome system significantly better than OGI system
- Bag of 3grams slightly ahead of Bigrams

2006 Performance

	BAG OF 3 GRAMS		BIGRAMS	
Test Type	**EER**	**MinDCF**	**EER**	**MinDCF**
1conv4w				
OGI	18.35	0.0668	18.68	0.0674
Callhome	15.54	0.0611	15.5	0.0613
3conv4w				
OGI	14.23	0.0559	15.33	0.0591
Callhome	11.62	0.0503	11.68	0.0519
8conv4w				
OGI	13.28	0.0503	14.73	0.0540
Callhome	9.87	0.0409	10.94	0.0426

Other notes:

- T-Norm and Z-Norm found to help significantly! ~ 15% relative improvement.
- Still need to go back and see how the individual streams performed. Consistent with other years?

Prosodic System

- N-gram modelling of prosodic events (similar to Adami's work †)
- Prosodic Event Labels:
  - Piecewise linear tokens describing joint pitch and energy trajectories. Categorised by:
    - Voiced or Unvoiced
    - Slope = [Rising, Falling]
    - Duration = [Short, Long]
  - Prosodic descriptions aligned with broad phonetic classes (same as used HMM system)
  - Total of 60 descriptive tokens.
- Modelling and Scoring:
  - Bag of 3grams and Bigrams
  - CT-Norm used

† Adami, Hermansky "Segmentation of Speech for Speaker and Language Recognition", Eurospeech 2003

Prosodic System Performance

	2006_DEV		2006	
	EER	MinDCF	EER	MinDCF
BAG-OF-3 (ALL)	21.79	0.0778	22.48	0.0813
BAG-OF-3 (ENG)	21.68	0.0772	21.35	0.0747
BIGRAM (ALL)	23.14	0.0776	23.06	0.0819
BIGRAM (ENG)	23.09	0.0764	21.72	0.0747

- Consistent performance in development and eval.
- Performance doesn't degrade considerably when non-English data added
- Z and T-Norm helped considerably. C-Norm didn't improve over Z-Norm.

Fusion

- Our fusion strategy was generally conservative
  - Reduce the risk of corpus mismatch issues
  - Reduce the potential for over-fitting given relatively high number of systems
  - Reduce the risk of human or scripting errors between Dev and Eval sets
    - But still not enough ☹
- Logisitic linear regression (LLR) was used to train a weighted-sum fusion with the FoCal package †
  - Simple…and calibration bonus…

† Niko Brümmer, "FoCal: Tools for Fusion and Calibration of automatic speaker detection systems," available at http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm, 2005

Fusion Metadata

- Transmission channel labels were added as metadata to the LLR fusion training motivated by the observations for the GMM system
  - The channel labels for the test segments were determined by **IBM's automatic channel detector**
- Gender labels were similarly added
- The metadata provided one of the biggest contributions to the fusion

Fusion Metadata

- The channel labels were encoded in 5 binary fields for each trial
  - For Carbon, CDMA, Electret, GSM and TDMA
  - Eg. [0 0 1 0 0] → Electret
- Effectively trains a bias for each channel type with LLR fusion
- A more aggressive strategy of training separate fusion systems for each channel type was rejected as too susceptible to over-fitting
  - This was backed by poor performance on the Eval set

DCF	Dev (2005)	Eval (2006)
LLR with Channel Biases	.0134	.0175
Separate Fusion by Channel	.0106	.0266

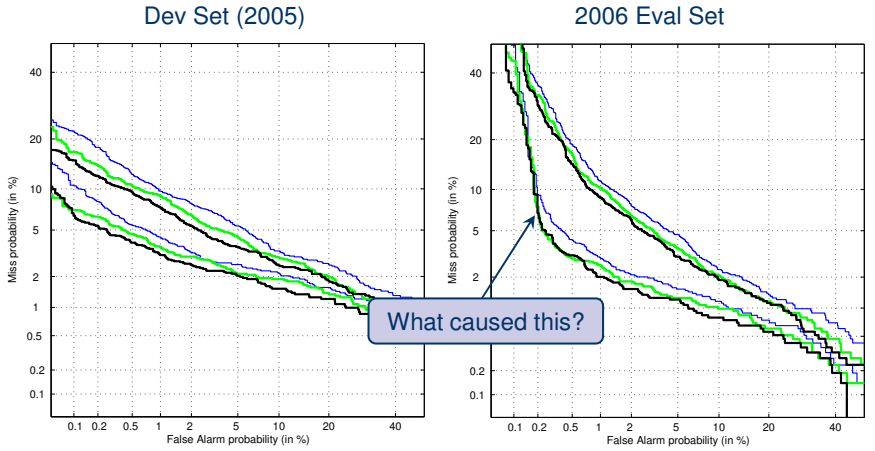
QUT Queensland University of Technology Brisbane Australia
a university for the real world
Speech and Audio Research Laboratory

Fusion Results

	Best Single System		QUT Only Fused		QUT + IBM Fused	
Test Set	EER	MinDCF	EER	MinDCF	EER	MinDCF
1conv4w						
Dev Set (2005)	5.26%	.0173	4.60%	.0148	3.89%	.0135
2006	4.69%	.0209	4.04%	.0188	3.89%	.0175
3conv4w						
Dev Set (2005)	3.48%	.0118	3.04%	.0102	2.84%	.0096
2006	2.25%	.0127	2.19%	.0118	2.12%	.0113
8conv4w						
Dev Set (2005)	2.84%	.0096	2.84%	.0080	2.40%	.0071
2006	2.10%	.0087	1.83%	.0068	1.64%	.0069

- The best single system was the forward Cepstral GMM system.
- The QUT Only fusion does not include channel metadata or the GMM Supervector SVM and Binary Tree systems. (all provided by IBM)
- The QUT + IBM fusion includes all available systems and metadata.

Fusion: 1- & 8-conv4w DET Plots



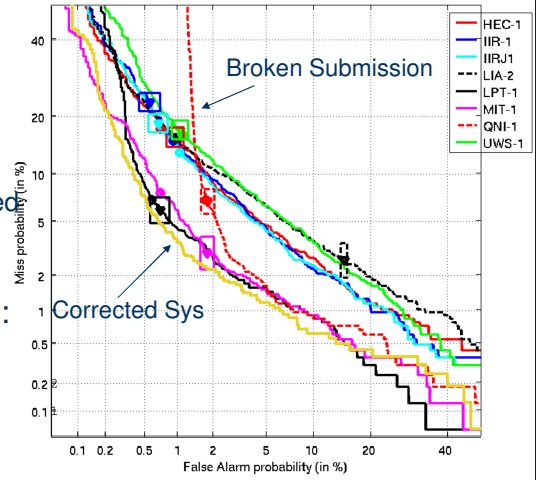
Single System **QUT Only Fusion** **QUT + IBM Fusion**

Fusion Outcomes

- Only modest gains due to fusion
  - Disappointing result
  - Adding channel metadata was helpful
  - The variety of systems included is apparently not sufficiently **diverse** / complementary
    - We need to investigate whether common elements in the front-end processing are to blame for this. Common front-end to all QUT and IBM systems!
- However, we believe the choice of conservative strategy was correct.

Corrected 3conv4w Submission

- An error was made in manual processing of 3conv4w QUT_1 submission.
  - **Handy Tip:** Make sure your score files are sorted before fusing!
- Corrected performance:
  - EER: 2.13%
  - MinDCF 0.0113



Some Conclusions…

- Overall, a fairly successful evaluation
  - Mainly due to channel factor GMM
  - Discussions/Collaboration with IBM very fruitful.
- Still not getting much out of fusion. (unlike other sites)
  - Diversity, diversity, diversity!
  - The SRE06 Fusion Group efforts may provide some answers!