NIST 2005
Speaker Recognition Evaluation
QUT Submission

Speaker: Robbie Vogt

Speech and Audio Research Laboratory
QUEENSLAND UNIVERSITY OF TECHNOLOGY

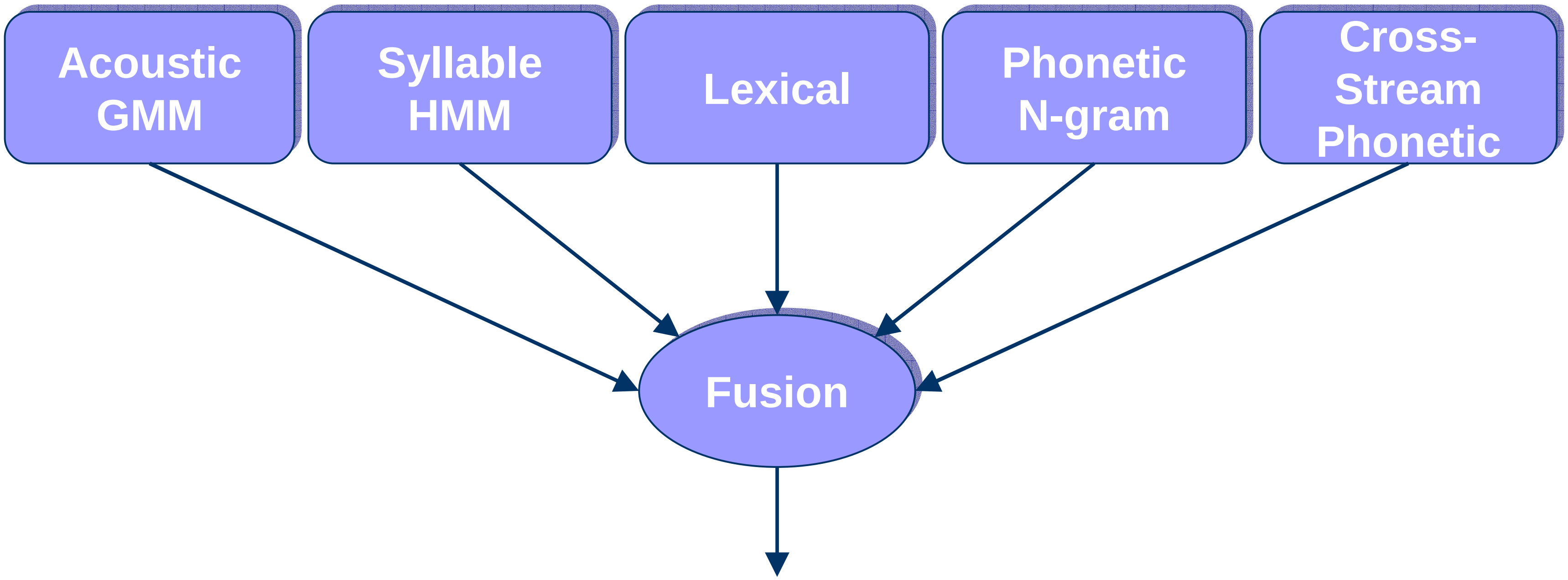
Robbie Vogt, Brendan Baker,
Michael Mason and Sridha Sridharan

Presentation Outline

- Submission Overview
- Development Data
- Acoustic Subsystem
  - Core GMM-UBM System
  - Channel Compensation
- Syllable-based HMM Subsystem
- Lexical Subsystem
- Phonetic N-gram Subsystem
- Cross-stream Phonetic Subsystem
- Fusion
- Overall System Performance

Submission Overview

- QUT_1 system comprises of 5 independent subsystems.



Submission Overview

- QUT_2 comprised of the acoustic-only system

- Evaluation conditions attempted
  - Results submitted for the 1 side testing and 1, 3 and 8 side training conditions.

Development Data

- NIST 2004 data used for most of the development data purposes
  - Background models
  - Individual system tuning
  - Fusion training
- Better matched conditions than other corpora, but limited in size and number of speakers

Development Data

- Switchboard-II data was also used when necessary
  - Significantly mismatched to Mixer data as demonstrated in SRE ’04
  - But lots of data and lots of speakers
  - Used to augment NIST 2004 data, not replace

Development Data

- A new evaluation protocol was developed using NIST 2004 data to overcome some of the limitations
  - Filtered out some of the less reliable sides from the original Evaluation protocol
    - Little or no data, erroneous speaker labels, etc
      - Removed 25 training and 3 test segments
    - We found these to have a significant effect on last year's results

Development Data

- A new evaluation protocol was developed using NIST 2004 data to overcome some of the limitations
  - 3 distinct splits with disjoint speaker sets
    - Similar to EDT protocols
    - Allowed for held-out set fusion development
  - ~300 models per split and ~45,000 trials
    - From ~100 speakers

Acoustic Subsystem

- Overview of acoustic system [1]
  - Feature warped MFCC features with appended delta [2]
  - GMM-UBM [3] based modelling and scoring with Channel Compensation based on [4]
  - Z-Norm and T-Norm [5].

Core GMM-UBM System

- Gender specific UBMs trained from pooled NIST 2004 and SWB-II data
  - 512 mixture components, 24-dimensional features
  - ~500 conversation sides for each gender, roughly half from SWB-II

Channel Compensation

- Channel variability (generally, session variability) was incorporated into the GMM modelling and scoring processes.
  - Based on [4] and similar in concept to the SDV submission last year.

Channel Compensation

- An utterance i is modelled by a GMM based on speaker and channel factors

$$\mathbf{m}_i(s) = \mathbf{m}(s) + \mathbf{U}\mathbf{x}_i(s)$$

  - Speaker is represented as a mean offset $\mathbf{m}(s)$ from the UBM independent of the channel
  - Channel is an additional mean offset $\mathbf{x}_i(s)$ restricted to 20-dimensional subspace $\mathbf{U}$

Channel Compensation: Enrolment

- During speaker enrolment, $\mathbf{m}(s)$ and all $\mathbf{x}_i(s)$ are optimised simultaneously
  - $\mathbf{m}(s)$ using classical MAP estimation, $\tau = 8$
  - $\mathbf{x}_i(s)$ using a MAP estimation with standard normal prior in the channel subspace
  - Only $\mathbf{m}(s)$ is retained

- Iterative approach used for the simultaneous optimisation
  - Similar to the Gauss-Seidel method

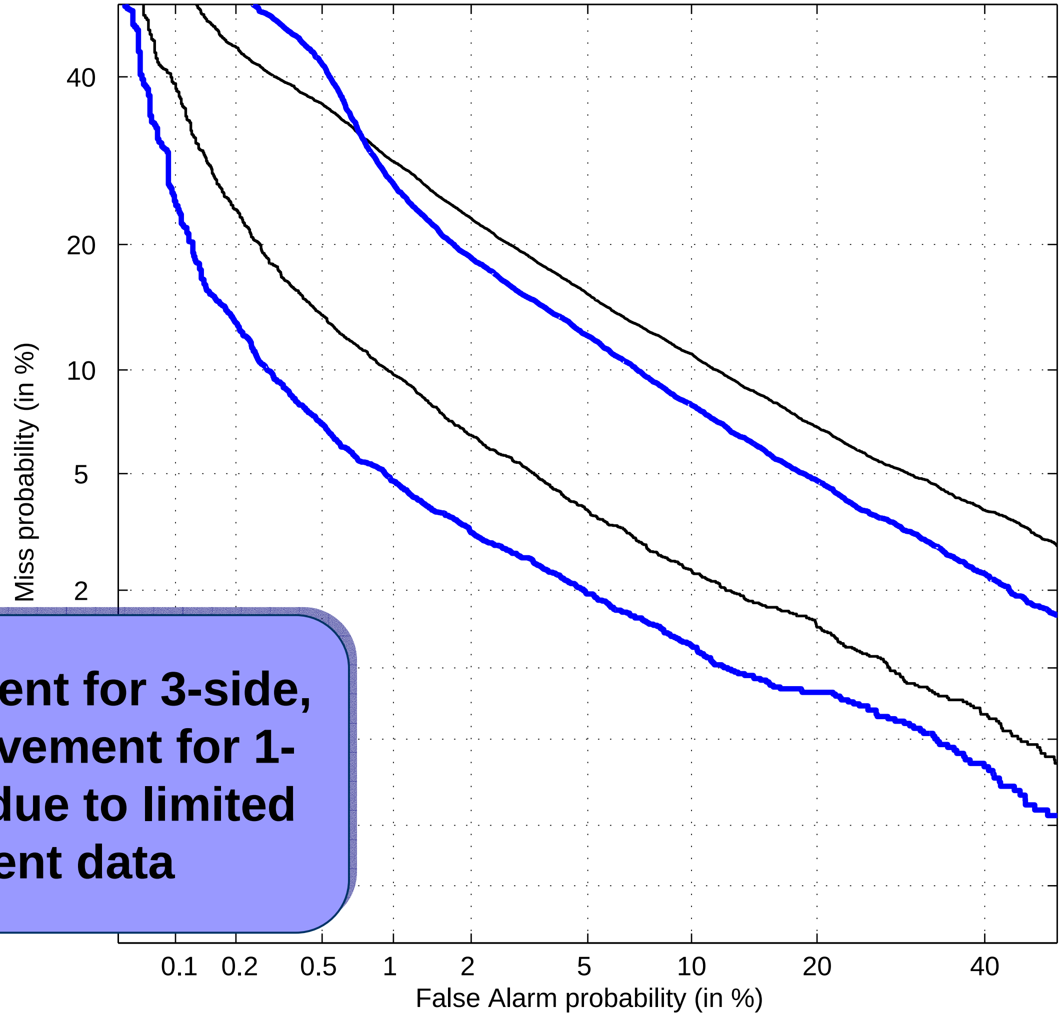
Channel Compensation: Scoring

- Essentially classical Top-N ELLR scoring, except
  - $\mathbf{x}_i(s)$ is estimated for each model / test segment combination first
  - This offset is applied to the speaker model means before scoring.
  - Some approximations are made for speed.

Channel Compensation Results

Comparison of baseline **ELLR** and **channel compensated** method for the development set.

1-side and 3-side training conditions.



Good improvement for 3-side, Reduced improvement for 1-side, probably due to limited development data

Normalisation

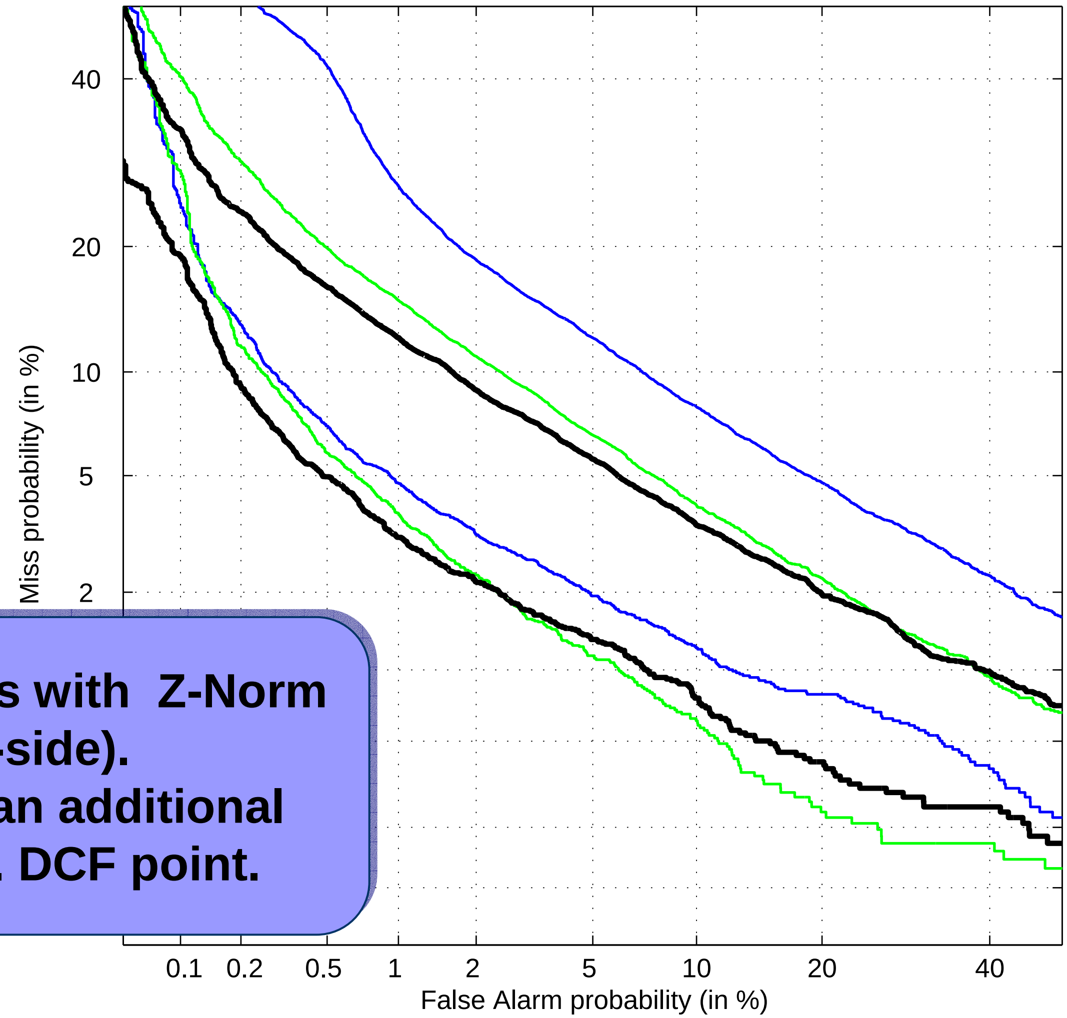
- Z-Norm segments selected from NIST 2004 data
  - From all 3 splits in our dev protocol for the evaluation
  - From the remaining 2 splits for each dev split
  - 260 total segments

- T-Norm models also from NIST 2004 data using distinct speakers
  - From 3 splits for the eval, 2 for dev
  - 200 total models for 1-side condition

Channel Compensation Results

Channel compensated method with **Z-Norm** and **ZT-Norm** for the development set.

1-side and 3-side training conditions.



Excellent results with Z-Norm (esp. 1-side).
T-Norm gave an additional boost at min. DCF point.

HMM Acoustic Subsystem using a "syllable"-length framework

- Very new work…Still under heavy development.
- Framework originally developed for language ID [6]
- Uses a pseudo-syllabic segmentation process. Modelling is then constrained to these segments.
- Allows for substitution of feature sets and modelling paradigms
- NIST2005 was the first attempt at using the framework with HMM modelling for speaker recognition.

Syllabic Segmentation

- Pseudo Syllabic Segmentation.
  - Multilingual broad phone recogniser used to recognise 4 phonetic classes
    - C1: Vowels/Dipthongs
    - C2: Nasals/Glides
    - C3: Fricatives
    - C4: Stops/Silences
  - Phone recogniser trained on OGI corpus – See [6] for further details.
  - Sliding window used to concatenate these broad phones into triplets forming our syllabic-like units.

Modelling & Feature Extraction

- A model is trained for each syllable resulting in 64 models in total.
- HMM topology used in the hope of capturing temporal information.
  - 7 state left-to-right HMM
  - 16 mixture components used for each emitting state
  - Speaker models adapted from appropriate background models using MAP.
- Feature extraction:
  - Same as the GMM system plus accelerations

Scoring

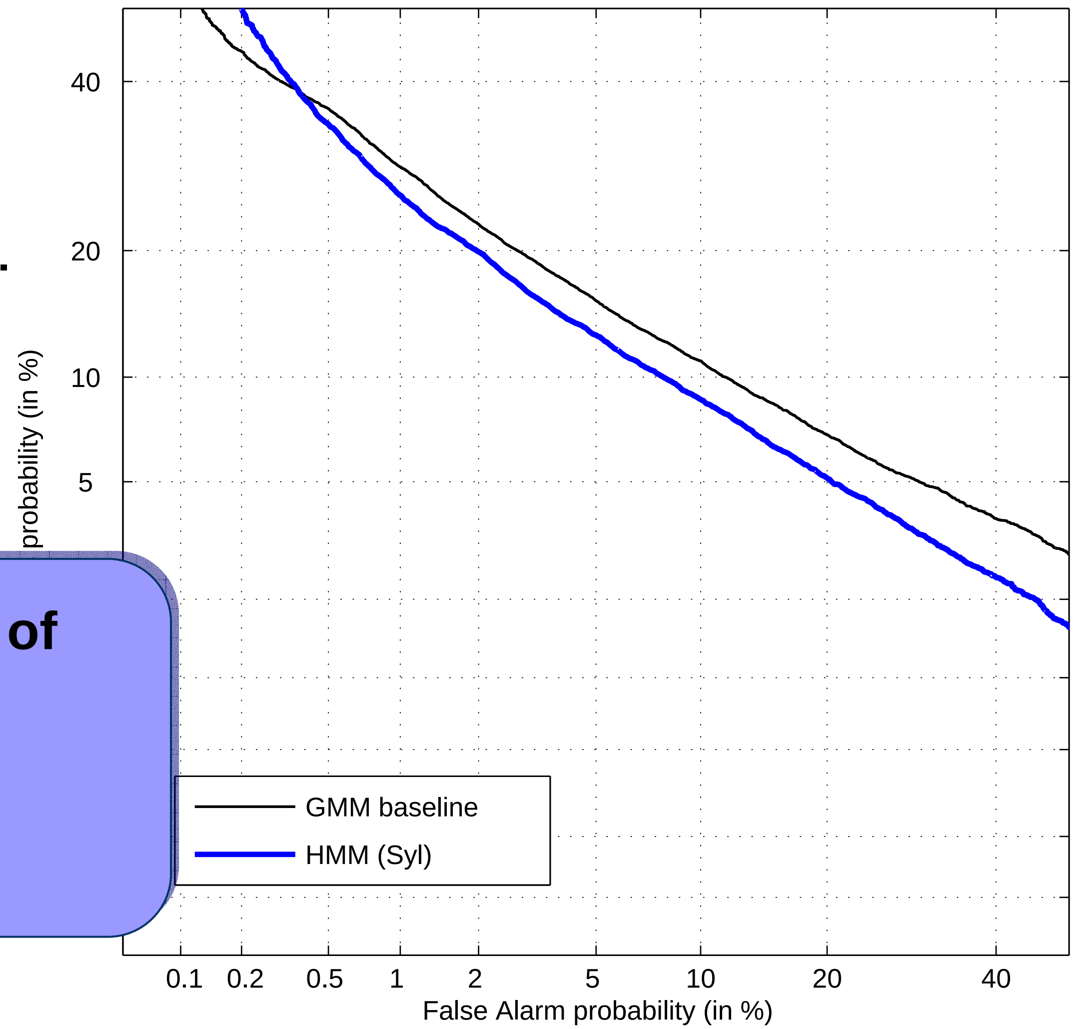
- System was developed so that there is a classifier for each syllable. System produces 64 scores for each test utterance.
- Only the **top 32** performing syllables (in terms of DCF) were used.
- Scores **fused** at output level using linear kernel SVM implemented in SVM Light
- No score normalisation performed due to time restrictions. (eg. T-Norm, Z-Norm)

HMM System Results

- Individual syllable classifiers produced EER in the range 13% - 45% on development data.
- Best performing syllable was *c2_c1_c2* (nasal/glide – vowel/dipthong – nasal/glide)
- Worst performing syllable was *c3_c3_c3* (Fricative – Fricative – Fricative).
- High correlation between rate of occurrence and performance.

HMM vs GMM Comparison

Comparison of development data results for baseline **GMM-UBM** and **HMM system** using syllable length framework.



HMM ahead for most of the curve.

MinDCF: GMM = 0.0389
HMM = 0.0356

HMM Future Work

- Initial results were pleasing.
- Since the evaluation, improvements have been made to phone recogniser. This may lead to better speaker rec performance.
- HMM configuration still to be optimised
  - Optimal # states
  - Mixture components
  - Adaptation factor
- Incorporation of score normalisation
  - T-Norm and Z-Norm

Lexical Subsystem

- Based on Doddington's word n-gram speaker recognition system. [7]
- Almost no change from QUT's 2004 lexical system.
- Modelling:
  - Bi-gram models used.
  - Target models adapted from UBM using n-gram MAP adaptation process.[8]
- UBM Source: Byblos ASR Transcriptions of NIST2004 SRE Data.

Phonetic N-gram Subsystem

- Utilises phone transcriptions from multiple Open-Loop Phone Recognisers (OLPR) each trained on 6 languages. Based on technique outlined in [9]
  - English, German, Hindi, Japanese, Mandarin, Spanish.
  - OLPR Trained on OGI corpus
- Very similar to last year's system
- Modelling :
  - Bag-of-N-grams. N=3 used
  - Target models adapted from UBM using n-gram MAP adaptation process.[8]
- This year we performed SAD before phone recognition. This helped a lot!

Phonetic N-gram System: Output

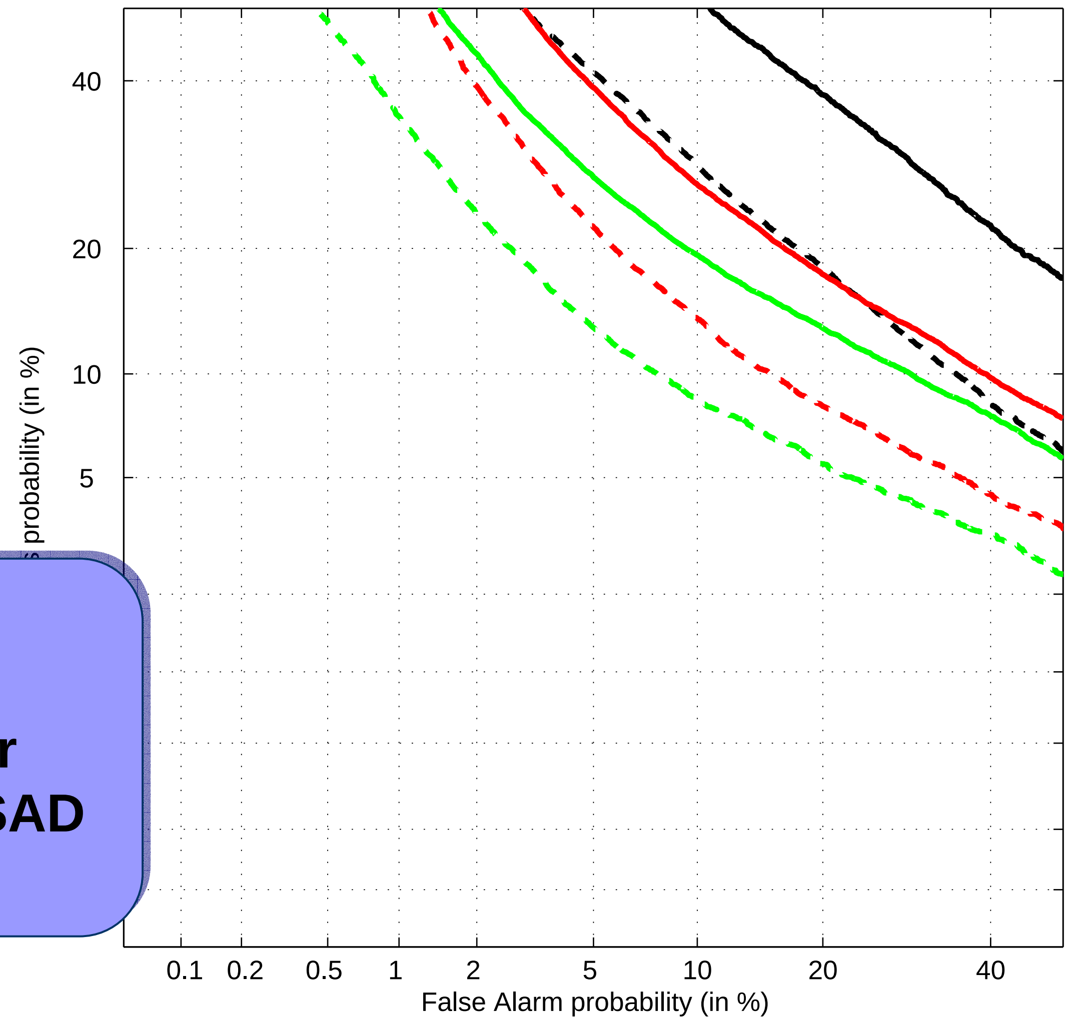
- Scores for each language stream were calculated using log-likelihood ratio.
- Scores for each language were then fused using a linear-kernel SVM (implemented with SVM Light)
- We found the SVM to be more stable than the MLP technique used last year.

Cross-stream Phonetic Subsystem

- Inspired by cross stream phonetic modelling performed by Jin et al. [10]
- Uses same phone streams as Phonetic N-gram System.
- Exploits patterns found across streams rather than in time dimension.
- Modeling:
  - Phone streams sampled every 15ms
  - The 6 phonetic events at each interval are used to form a token.
  - Unigram modelling of these tokens was performed.
  - MAP adaptation used to combat model sparsity
  - Pruning threshold also required to reduce model size.

High-level Feature Performance

Results on development data for the **Lexical** , **Phonetic N-gram** and **Cross-Stream Phonetic** features for 1side (solid) and 3side (dashed) conditions.



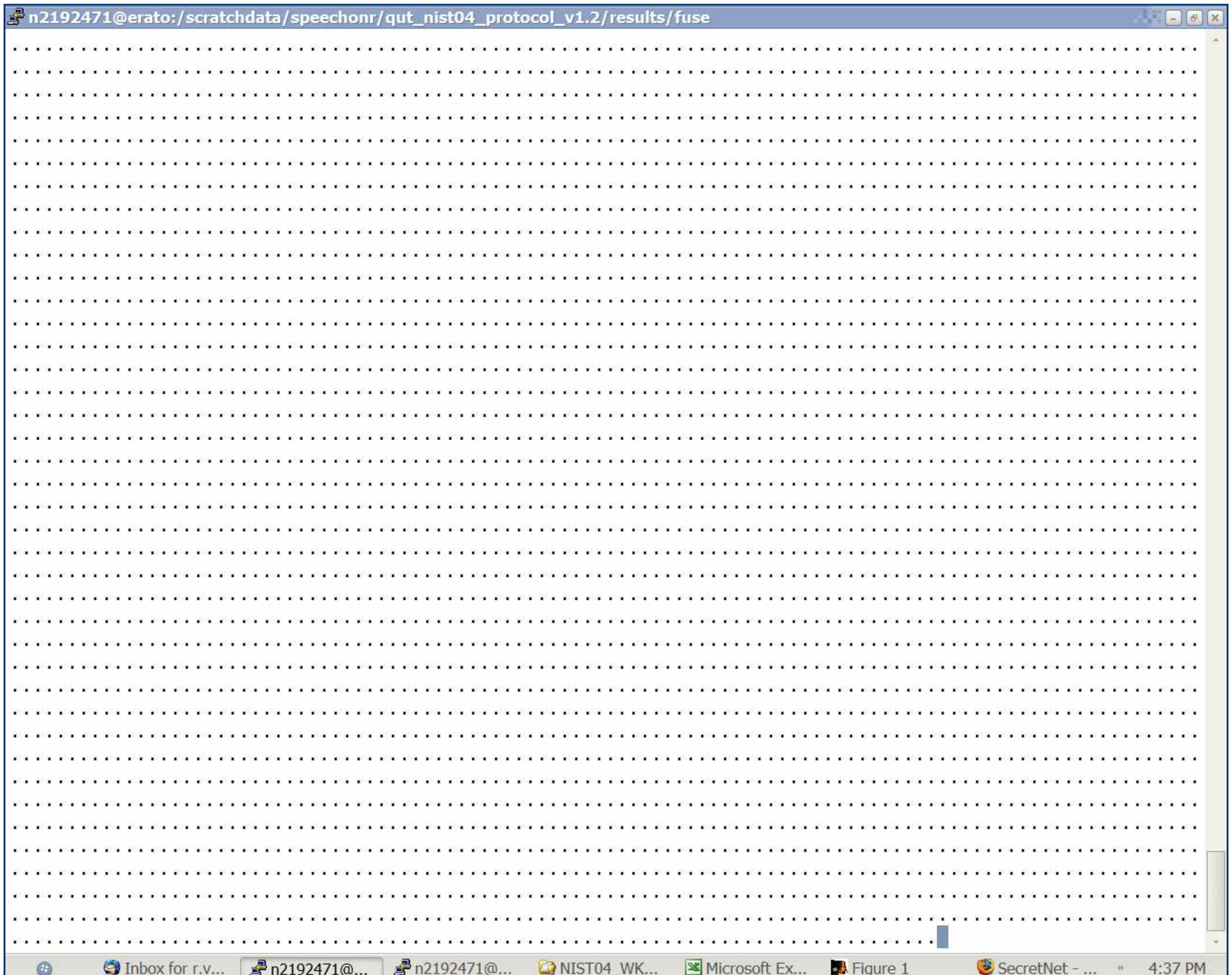
Results as expected.

Phonetic results better than last year due to SAD process.

Fusion

- Last year we used an MLP to fuse our subsystems
  - Issues with corpus mismatch and operating point stability
- This year we used SVMs implemented with SVM Light.

Fusion


n2192471@erato:/scratchdata/speechonr/qut_nist04_protocol_v1.2/results/fuse
Inbox for r.v...
n2192471@...
n2192471@...
NIST04_WK...
Microsoft Ex...
Figure 1
SecretNet - ...
4:37 PM

Fusion: Some Details…

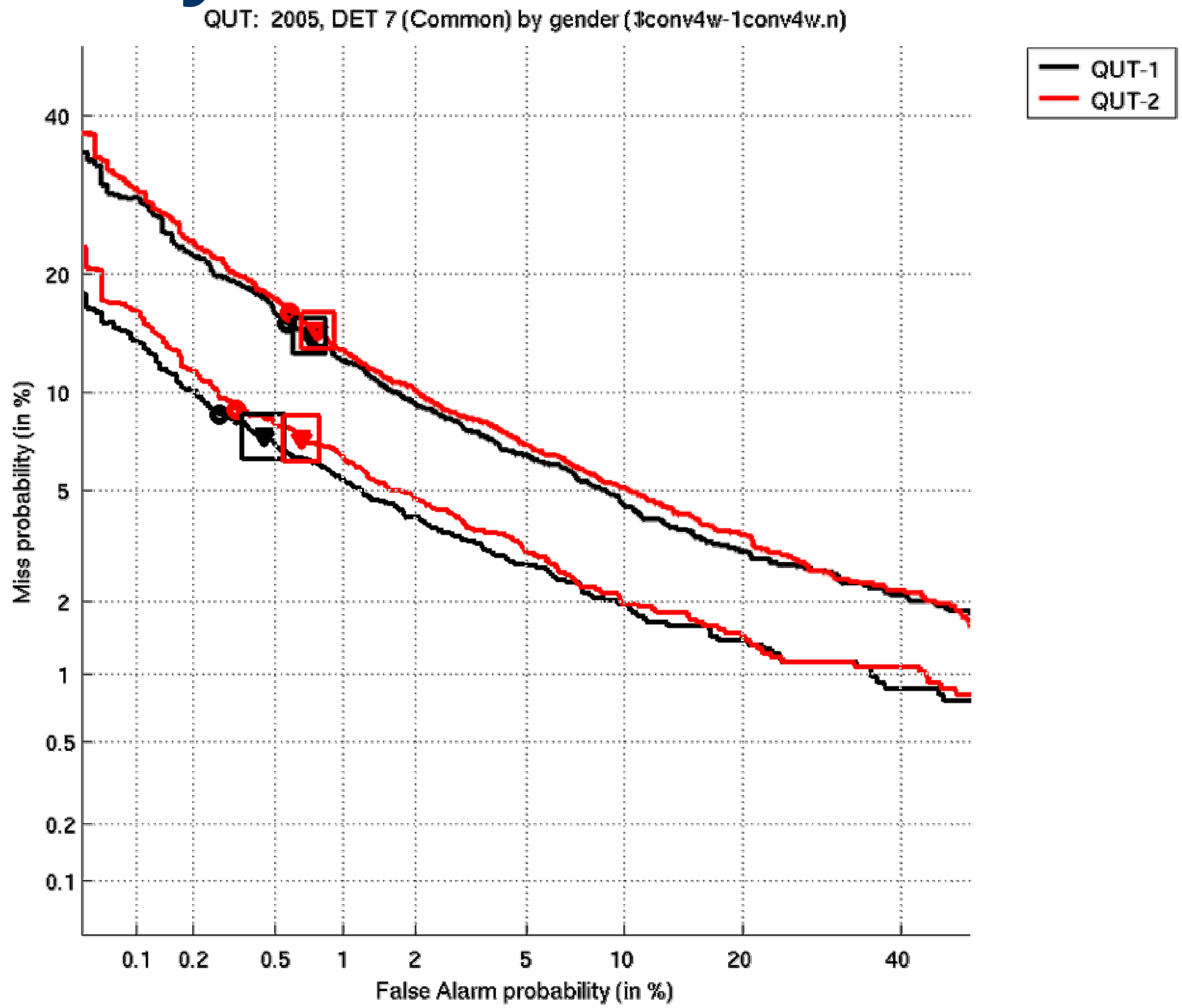
- An SVM was trained on each split of the development data
  - Linear kernel to avoid mismatch / stability issues
  - Final result was the averaged result from the 3 SVM classifiers

- Inputs:
  - Acoustic, Syllable HMM, Phonetic, Cross Stream, Lexical + Gender

Fusion: Conclusions

- Only small gains in performance…
- …but the fusion had a more stable operating point.

Overall System Performance



References (i)

[1] R. Vogt, B. Baker, and S. Sridharan, "Modelling Session Variability in Text-Independent Speaker Verification," in proc. Interspeech, 2005, submitted.

[2] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," In Proc. of A Speaker Odyssey, pp. 213-218, 2001.

[3] D. A. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 19-41, January 2000.

[4] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in proc.Odyssey: The Speaker and Language Recognition Workshop, 2004.

[5] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, "Score Normalisation for Text-Independent Speaker Verification Systems," Digital Signal Processing, vol. 10, pp. 42-54, January 2000.

References (ii)

[6] T. Martin, B. Baker, E. Wong, and S. Sridharan, “A syllable-length framework for language identification,” In print, *Computer Speech and Language*, 2005.

[7] G. Doddington, "Some Experiments on Ideolectal Differences Among Speakers," http://www.nist.gov/speech/tests/spk/2001/doc/n-gram_experiments-v06.pdf, 14 November 2000.

[8] B. Baker, R. Vogt, M. Mason, and S. Sridharan, "Improved Phonetic and Lexical Speaker Recognition through MAP Adaptation," in proc.Odyssey: The Speaker and Language Recognition Workshop, 2004.

[9] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 149 - 152, 2002.

[10] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews and J. Abramson, “Combining cross-stream and time dimension in phonetic speaker recognition" in proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 800-803, 2003.