Institute for
Infocomm Research

NIST SRE 2010 I4U Consortium

(IIR, USTC/iFly, UEF, UNSW, NTU)

Presented by Bin MA and Kong Aik LEE

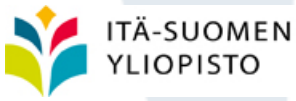
Brno, Czech Republic
June, 2010

Agenda

- The I4U Consortium
- System Overview
- Subsystems & Features
- Fusion Strategies & Threshold Setting
- Development Dataset Design
- Analysis of Results
- Conclusions

The I4U Consortium

- IIR + Researchers and Interns from Universities
  - Institute for Infocomm Research (**I**IR), Singapore
  - University of Science and Technology of China (**U**STC/iFly), China
  - University of Eastern Finland (**U**EF), Finland
  - University of New South Wales (**U**NSW), Australia
  - Nanyang Technological University (NT**U**), Singapore



The I4U Consortium

- **IIR**
  - Haizhou Li, Bin Ma, Hanwu Sun, Kong Aik Lee, Changhuai You, Donglai Zhu, Chien-Lin Huang, Cheung Chi Leung, Ville Hautamaki
  - Hosting Organization
  - Focus: Dev set design, Front-end, Baseline speaker classifiers
- **USTC/iFly**
  - Wu Guo, Eryu Wang, Lirong Dai
  - Focus: Joint Factor Analysis for GMM-UBM, Front-end
- **UEF**
  - Tomi Kinnunen, Rahim Saeidi, Filip Sedlak, Pasi Franti
  - Focus: Stabilized Weighted Linear Prediction (SWLP)
- **UNSW**
  - Karen Kua, Thiruvaran Tharmarajah, Mohaddeseh Nosratighods, Eliathamby Ambikairajah, Julien Epps
  - Focus: Spectral Centroid Frequency (SCF) and Spectral Centroid Magnitude (SCM)
- **NTU**
  - Rong Tong, Eng Siong Chng
  - Focus: Fusion

NIST SRE10 – Post-mortem

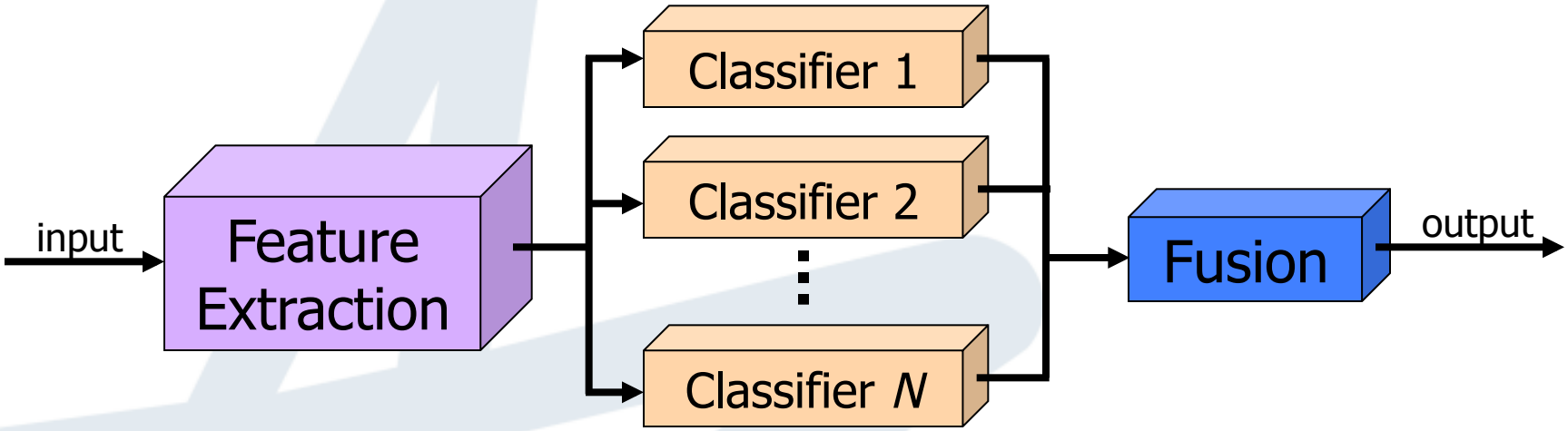
- I4U – joint team from strength to strength.
- Carefully designed development set and baseline speaker classifiers shared across members.
- Brute-force search of fusion weights minimizing the Detection Cost Function (DCF).

- Fusion and decision thresholding by channel type and gender can be improved.
- Engineering pitfall – score calibration has to be included for the case where the fusion and threshold are done in a gender-dependent manner.

Agenda

- The I4U Consortium
- System Overview
- Subsystems & Features
- Fusion Strategies & Threshold Setting
- Development Dataset Design
- Analysis of Results
- Conclusions

System overview



- Architecture:
  - Fusion of multiple classifiers
- Three major components:
  - Feature extraction
  - Parallel classifiers
  - Linear score fusion

Classifiers and features

Classifier	Feature
GMM-UBM-JFA	LPCC
GMM-SVM-KL	PLP
GMM-SVM-BHATT*	MFCC
GMM-SVM-FT*	SCM-SCF*
	SWLP*

Note:

- (*) indicates new efforts
- JFA – Joint Factor Analysis
- KL – Kullback-Leibler divergence
- BHATT – Bhattacharyya distance
- FT – Feature transformation
- SCM – Spectral centroid magnitude
- SCF – Spectral centroid frequency
- SWLP – Stabilized weighted linear prediction

New efforts

- Classifiers
  - Feature Transformation (FT) based GMM-SVM [1]
  - Bhattacharyya Kernel based GMM-SVM [2]
- Features
  - Spectral Centroid Magnitude – Spectral Centroid Frequency (SCM-SCF) [3]
  - Stabilized Weighted Linear Prediction (SWLP) [4]

---

[1] D. Zhu, B. Ma and H. Li, "Joint MAP adaptation of feature transformation and Gaussian mixture model for speaker recognition," in *Proc. ICASSP*, 2009.

[2] C. You, K. A. Lee, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Trans. Audio, Speech, and Language Processing*, in press.

[3] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition", to appear in *Odyssey Speaker and Language Recognition Workshop*, 2010.

[4] R. Saeidi, J. Pohjalainen, T. Kinnunen, P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification", *IEEE Signal Processing Letters*, in press.

Agenda

- The I4U Consortium
- System Overview
- Subsystems & Features
- Fusion Strategies & Threshold Setting
- Development Dataset Design
- Analysis of Results
- Conclusions

Subsystems and Features

- The I4U system employs four classification techniques in combination with five different acoustic features.
- 13 subsystems were developed based on similar set of development data.
- Serves as useful resources for future study of score calibration and fusion methods.

Classifier	Feature
GMM-UBM-JFA	LPCC
GMM-SVM-KL	PLP
GMM-SVM-BHATT (NEW)	MFCC
GMM-SVM-FT (NEW)	SCM-SCF (NEW)
	SWLP (NEW)

Classifier 1 of 4: GMM-SVM

- Implemented based on:

  [1] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc*. ICASSP, pp. 97-100, 2006.

  [2] D. A. Reynolds, T. F. Quatieri and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Process., vol. 10, 19–41, 2000.

- MAP adaptation from UBM (with 512 mixtures) to obtain the speaker-dependent GMM.
- Form GMM supervector by stacking the mean vectors.
- NAP (nuisance attribute projection) removes the subspaces (rank 60) related to channel variability.
- Linear kernel SVM (KL divergence)

$$\kappa_{\mathrm{KL}}\left(\mathrm{X}_a,\mathrm{X}_b\right)=\sum_{i=1}^{N}\lambda_i\mathbf{m}_i^{(a)}\boldsymbol{\Sigma}_i^{-1}\mathbf{m}_i^{(b)}=\sum_{i=1}^{N}\left(\sqrt{\lambda_i}\boldsymbol{\Sigma}_i^{-1/2}\mathbf{m}_i^{(a)}\right)^T\left(\sqrt{\lambda_i}\boldsymbol{\Sigma}_i^{-1/2}\mathbf{m}_i^{(b)}\right)$$

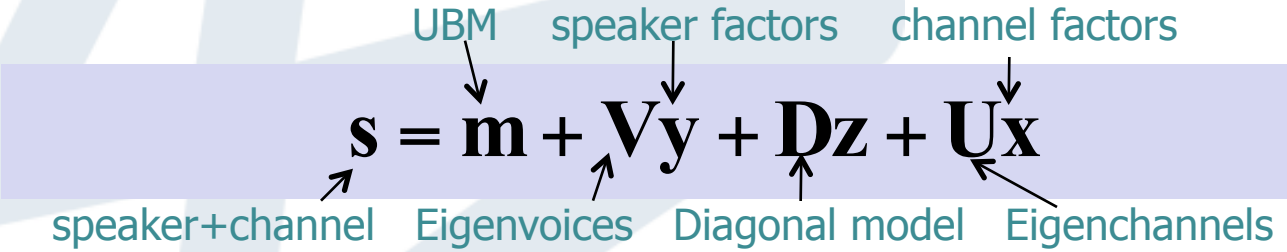
- Pre-computed kernel matrix for fast training of speaker models.
- Model compaction for fast scoring.
- T-norm was used for score normalization.

Classifier 2 of 4: Joint Factor Analysis (JFA) (1/2)

- Implemented based on:

  P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," IEEE Trans. on Audio, Speech and Language Processing, July 2008.

- We used standard JFA model:



	UBM	V	D	U
Train Data	SRE04 tel	SW II tel SRE04 tel	SRE04 tel	SRE04, SRE05, SRE06, SRE08, Mixer5, SRE08 follow-up
Config.	1024	300	Diagonal	TEL (100) ITV (100) MIC (50)

Classifier 2 of 4: Joint Factor Analysis (JFA) (2/2)

- Train **V**, then **U**, and finally **D**. For **U**, **U**_tel, **U**_mic and **U**_itv were trained separately and combined.
- Linear scoring was used, ref:

  O. Glembek, L. Burget, N. Dehak, N Brummer, P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in Proc. ICASSP, pp. 4057 – 4060, 2009.
- Score normalization:
  - SRE05 (tel and mic) and SRE06 (tel and mic).
  - T-norm follows the channel type of the training segment.
  - Z-norm follows the channel type of the test segment.

Classifier 3 of 4: FT-SVM (1/2)

- Given a speaker-dependent utterance X, a transformation function F is learned in order to produce Y which is speaker-independent:

$$\mathbf{y}_t = \mathrm{F}\left(\mathbf{x}_t;\Theta\right) = \mathbf{A}_k\mathbf{x}_t + \mathbf{b}_l, \quad \mathrm{X} = \left\{x_t\right\}_{t=1}^{T}, \mathrm{Y} = \left\{y_t\right\}_{t=1}^{T}$$

transformation matrix bias

- The transformation matrices $\{\mathbf{A}_k: k = 1,\ldots,K\}$ and biases $\{\mathbf{b}_l: l = 1,\ldots,L\}$ are determined such that the transformed utterance Y is as similar to the UBM (speaker-independent) by maximizing the following objective function:

$$L\left(\Theta,\Lambda\right) = p\left\{\mathrm{F}\left(\mathbf{x}_t;\Theta\right)\middle|\Lambda\right\}\prod_{k=1}^{K}p\left(\mathbf{A}_k\right)\prod_{l=1}^{L}p\left(\mathbf{b}_l\right)$$

feature transformation UBM prior density

- Use the parameters (i.e., the matrices and bias vectors) of the feature transformation (FT) function to form the supervector.

Classifier 3 of 4: FT-SVM (2/2)

- We use UBM with 512 mixtures, K = 1 and L = 512. This produces 1 transformation matrix and 512 bias vectors used to form the supervector.
- In GMM-SVM and MLLR-SVM, the adaptation or transformation is applied on the model parameters. Here, the transformation is applied on the feature vectors.
- Compared to CMLLR (feature based), FT is different in the following aspects [1]:
  - Transformation matrices and bias vectors are assigned to regression classes separately (more flexible).
  - Uses the MAP criterion in the estimation process.
    - Avoid probably numerical problems in ML methods (e.g., CMLLR) caused by insufficient training data.

- Ref: D. Zhu, B. Ma and H. Li, "Joint MAP adaptation of feature transformation and Gaussian mixture model for speaker recognition," in *Proc. ICASSP*, 2009.

Classifier 4 of 4: GMM-SVM using Bhattacharyya Dist.

- Use Bhattacharyya distance instead of KL divergence as the distance measure between GMM supervectors.
- In addition to the mean vectors, covariance matrices are adapted as well via MAP.
- The Bhattacharyya kernel is given by:

$$\kappa_{\mathrm{BHATT}}\left(\mathrm{X}_{\mathrm{a}},\mathrm{X}_{\mathrm{b}}\right)=\sum_{i=1}^{N}\left[\left(\frac{\boldsymbol{\Sigma}_i^{(\mathrm{a})}+\boldsymbol{\Sigma}_i^{\mathrm{UBM}}}{2}\right)^{-1/2}\left(\mathbf{m}_i^{(\mathrm{a})}-\mathbf{m}_i^{(\mathrm{UBM})}\right)\right]^{T}\left[\left(\frac{\boldsymbol{\Sigma}_i^{(\mathrm{b})}+\boldsymbol{\Sigma}_i^{\mathrm{UBM}}}{2}\right)^{-1/2}\left(\mathbf{m}_i^{(\mathrm{b})}-\mathbf{m}_i^{(\mathrm{UBM})}\right)\right]$$

- Differences from the KL kernel:
  - Covariance matrices are adapted and appear as normalization factor. Mixture weights are not part of the kernel.
  - UBM supervector is part of the kernel. Since the covariance matrices are adapted, this introduces different shifting to supervectors.

$$\kappa_{\mathrm{KL}}\left(\mathrm{X}_{a},\mathrm{X}_{b}\right)=\sum_{i=1}^{N}\left(\sqrt{\lambda_i}\boldsymbol{\Sigma}_i^{-1/2}\mathbf{m}_i^{(a)}\right)^{T}\left(\sqrt{\lambda_i}\boldsymbol{\Sigma}_i^{-1/2}\mathbf{m}_i^{(b)}\right)$$

- Ref: C. You, K. A. Lee, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Trans. Audio, Speech, and Language Processing*, in press.

Feature 1 of 2: Spectral centroid frequency + magnitude

- The spectrum $S(f)$ of each speech frame is partitioned into subbands (lower and upper cut-off frequencies given by l_k and u_k).
- The frequency F_k and magnitude M_k of the subband centroid are computed, as follows:

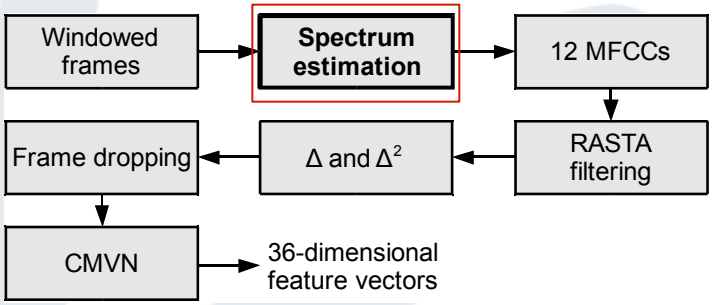
$$F_k = \frac{\sum_{f=l_k}^{u_k} f\left|S(f)w_k(f)\right|}{\sum_{f=l_k}^{u_k}\left|S(f)w_k(f)\right|} \qquad M_k = \frac{\sum_{f=l_k}^{u_k} f\left|S(f)w_k(f)\right|}{\sum_{f=l_k}^{u_k} f}$$

- The weights are given by the normalized energy of frequency points in that subband.
- Each subband produces two features. For K = 14 subbands, each frame produces 28 dimensional SCF-SCM feature vector.
- SCF and SCM capture formant related information.
- Contributed by UNSW, for details please see:

J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition", to appear in *Odyssey Speaker and Language Recognition Workshop*, 2010.

Feature 2 of 2: Stabilized weighted linear prediction (SWLP)

- Similar to conventional MFCC except that the FFT spectrum is replaced by an all-pole spectrum obtained through a stabilized weighted linear prediction (SWLP).



- The idea of SWLP analysis is to weight the LP residual with a short-term energy function so that the spectrum estimation focuses on high-energy portion (i.e., less corrupted by additive noise) within a speech frame.
- Contributed by UEF, for details please see

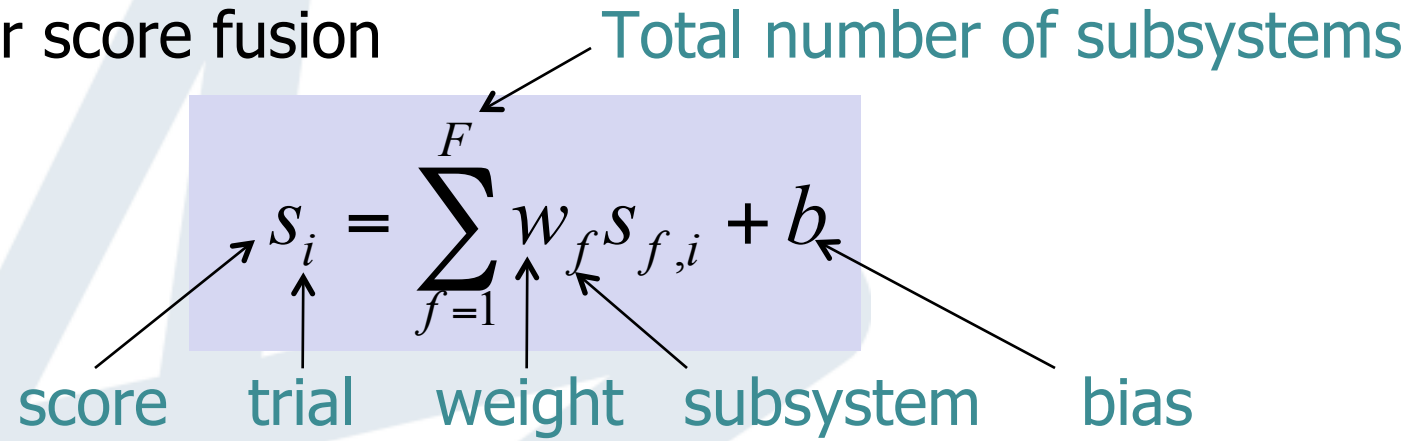
R. Saeidi, J. Pohjalainen, T. Kinnunen, P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification", *IEEE Signal Processing Letters*, in press.

Agenda

- The I4U Consortium
- System Overview
- Subsystems & Features
- Fusion Strategies & Threshold Setting
- Development Dataset Design
- Analysis of Results
- Conclusions

Fusion strategy

- Linear score fusion



- Brute-force search of fusion weights by minimizing the min DCF.
- Gender-dependent fusion and threshold (no-cross gender trials).
- Threshold, Θ , is given by the min DCF point derived from the dev set.
- Scores are shifted based on the threshold, $b = -\Theta$.

Brute-force search of fusion weights

- Tune fusion weights:

$$\left(\hat{w}_f, \hat{b}\right) = \arg\min_{w_f, b} DCF\left(\sum_{f=1}^{F} w_f s_{f,i} + b\right)$$

- Adjust w_f one by one iteratively (a flat start was used).
- At each iteration, the weight w_f is determined with a simple grid search between $(w_f - \Delta)$ and $(w_f + \Delta)$ with a step size of $\Delta\Delta$, where $\Delta = 0.1$ and $\Delta\Delta = 0.01$.
- The optimum weight is the one that produces the min DCF value within the search window.
- The fusion weights are sum to 1 before proceed to the next iteration.

Agenda

- The I4U Consortium
- System Overview
- Subsystems & Features
- Fusion Strategies & Threshold Setting
- Development Dataset Design
- Analysis of Results
- Conclusions

SRE10 CORE task

- Consists of 9 subtasks
- Two **styles** of speech:
  - interview
  - telephone conversational
- Two types of **channel**:
  - microphone
  - telephone
- This ends up with three **channel conditions** to deal with
  - **ITV** (interview style, microphone channel)
  - **TEL** (telephone conversational, telephone channel)
  - **MIC** (telephone conversational, microphone channel)

Overview of SRE10 CORE task

Train-Test	Subtasks
ITV-ITV	1. Same microphone in train and test 2. Different microphone in train and test
ITV-TEL	3. Normal vocal effort
ITV-MIC	4. Normal vocal effort
TEL-TEL	5. Normal vocal effort 6. Normal in train, high in test 8. Normal in train, low in test
TEL-MIC	7. Normal in train, high in test 9. Normal in train, low in test

- We group the nine subtasks into five categories according to the training and test channels.
- Vocal effort (high or low) not handled.
- ITV-MIC train-test condition is new. No data available from previous SREs to simulate this condition.
- We used the fusion weights trained on ITV-ITV for the new ITV-MIC condition.

Development data

- Development data for classifier
  - UBM training, SVM background
  - Channel compensation (eigenchannel, NAP)
  - Eigenvoice modeling
  - Score normalization (t-norm, z-norm, tz-norm, zt-norm)
- Development data for fusion and threshold setting.

Development data	Description	Used for
SRE04	TEL	Classifier
SRE05	TEL, MIC (8-microphone configuration)	Classifier
SRE06	TEL, MIC (8-microphone configuration)	Classifier
Mixer5 (6 speakers)	ITV (16-microphone configuration)	Classifier
SRE08 + (followup data)	TEL, MIC (16-microphone configuration), ITV (16-microphone configuration) 844 female + 492 male speakers	Classifier (40%) Fusion (60%)
Switchboard II	Cell, landline	Eigenvoice modeling

Development set design

Train-Test	Number of models	Number of trials
itv-itv	900 (520 f + 380 m)	845911
itv-tel	900 (520 f + 380 m)	91665
tel-tel	2027 (1290 f + 737 m)	341000
mic-mic	252 (135 f + 117 m)	67100
itv-mic	Not available	Not available
tel-mic	Not required	Not required
mic-tel	Not required	Not required

- Development set was constructed using 60% of the data from SRE08 and SRE08-followup.
- We used different sets of fusion weights for each subtasks and gender.
- The probability of target was set to 0.01 (i.e., one true speaker in 100 trials) in designing the development set.

Development set vs SRE10 evaluation set

DEV set Subtasks	SRE10 EVAL Set Subtasks	Ptarget (SRE10)
ITV-ITV	1. Same microphone in train and test 2. Different microphone in train and test	0.034 0.034
ITV-TEL	3. Normal vocal effort	0.028
ITV-MIC	4. Normal vocal effort	0.028
TEL-TEL	5. Normal vocal effort 6. Normal in train, high in test 8. Normal in train, low in test	0.023 0.013 0.013
TEL-MIC	7. Normal in train, high in test 9. Normal in train, low in test	0.010 0.011

- For SRE10 evaluation data the probability of target for the 9 subtasks was found to be in the range 0.010 ~ 0.034 (we assumed 0.01 in our development set).
- The number of trials are in the same order of magnitude with our development set.

Agenda

- The I4U Consortium
- Subsystems & Features
- Fusion Strategies & Threshold Setting
- Development Dataset Design
- Analysis of Results
- Conclusions

Average classifier performance on SRE10 (1/3)

Subsystems	SRE10 EVAL	
	Average EER (%)	Average MinDCF
JFA (PLP)	3.549126	0.052402
GMM-SVM (MFCC)	3.585587	0.052838
USTC-JFA (PLP)	3.631718	0.063529
GMM-SVM-BHATT (PLP)	3.727385	0.055580
JFA II (PLP)	3.878684	0.059104
USTC-SVM (PLP)	4.303088	0.054708
GMM-SVM (LPCC)	4.620497	0.067774
GMM-SVM (MLF)	4.784512	0.064650
FT-SVM (PLP)	5.286211	0.063744
USTC-SVM (LPCC)	5.506954	0.064260
SWLP	6.100724	0.067034
USTC-JFA (LPCC)	6.493085	0.077431
SCM-SCF	7.267840	0.067483

- JFA gives the best average performance in terms of EER and MinDCF.
- The new methods GMM-SVM-BHATT and FT-SVM give satisfactory performance. Further improvement required for SWLP and SCM-SCF.

Average classifier performance on SRE10 (2/3)

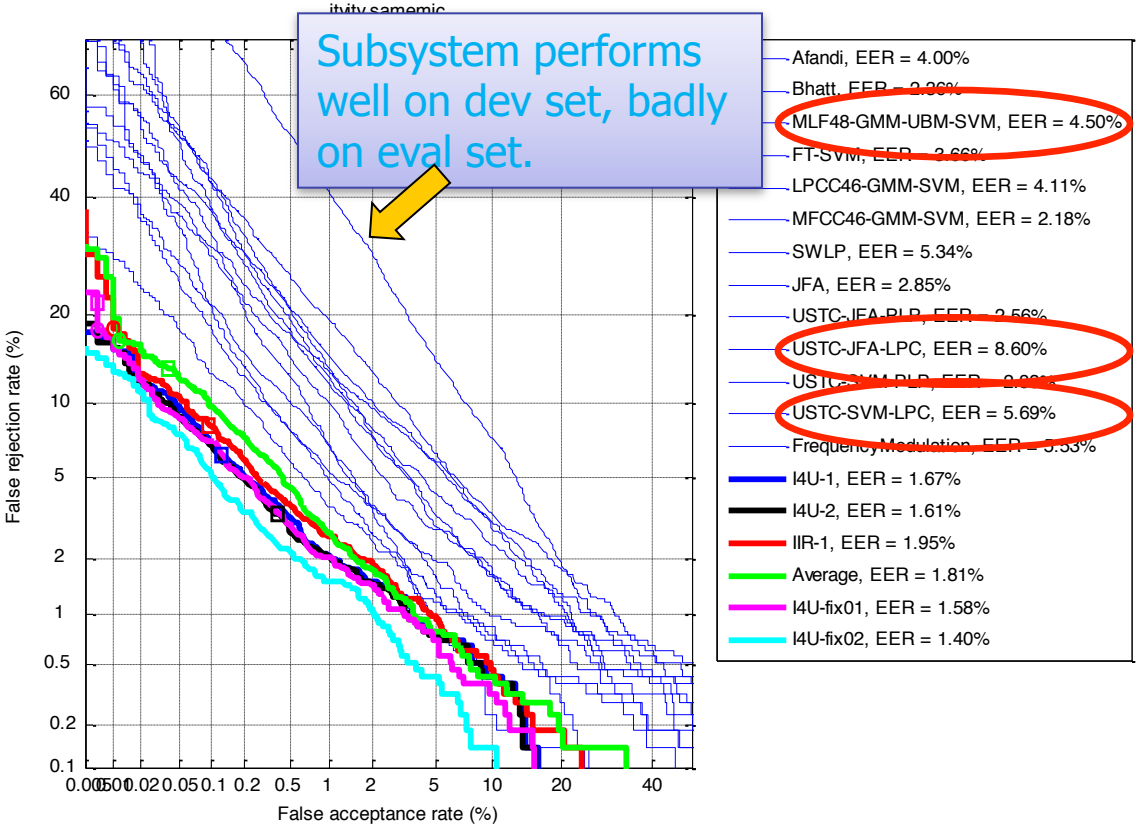
- Performance of subsystems ranked according to EER (averaged across subtasks):

SRE10 EVAL	EER
JFA (PLP)	3.549126
GMM-SVM (MFCC)	3.585587
USTC-JFA (PLP)	3.631718
GMM-SVM-BHATT (PLP)	3.727385
JFA II (PLP)	3.878684
USTC-SVM (PLP)	4.303088
GMM-SVM (LPCC)	4.620497
GMM-SVM (MLF)	4.784512
FT-SVM (PLP)	5.286211
USTC-SVM (LPCC)	5.506954
SWLP	6.100724
USTC-JFA (LPCC)	6.493085
SCM-SCF	7.267840

DEV Set	EER
GMM-SVM (MLF)	3.095955
JFA II (PLP)	3.119711
USTC-JFA (PLP)	3.223851
JFA (PLP)	3.244834
GMM-SVM (MFCC)	3.351505
GMM-SVM-BHATT	3.404958
USTC-JFA (LPCC)	3.541957
USTC-SVM (PLP)	3.808943
USTC-SVM (LPCC)	3.992951
GMM-SVM (LPCC)	4.083766
FT-SVM (PLP)	4.591922
SWLP	6.014758
SCM--SCF	6.370079

- Dividing all the subsystems into three tiers, it can be seen that most subsystems (10 out of 13) exhibit consistent performance in the DEV and SRE10 EVAL sets.

Average classifier performance on SRE10 (3/3)



- There are three subsystems that show inconsistent performance:
  - GMM-SVM (MLF) – over-trained on the DEV set.
  - LPCC (JFA and SVM) – front-end over-tuned on DEV set.

Fusion strategy – post evaluation (1/2)

- Gender-dependent fusion creates two separate thresholds for male and female scores.
- Shifting the scores according to the thresholds causes the male and female score distributions to be shifted differently.
- The actual DCF is not affected. But the DET curve, MinDCF, and EER are greatly affected.
- Modification:
  - STEP 1: Calibrate the score of individual susbsystems prior to fusion (FoCal toolkit was used)
  - STEP 2: Brute-force search of fusion weights

Fusion strategy – post evaluation (2/2)

– STEP 3: Threshold set according to the cost:

$$\theta = -\text{logit}\left(P_{\text{target}}\right) - \log\left(\frac{C_{\text{miss}}}{C_{\text{fa}}}\right)$$

i.e., one threshold for male and female scores instead of two separate thresholds. Scores from individual subsystems have been calibrated and the fusion weights sum to one.

Cmiss	Cfa	Ptarget	θ
1.0	1.0	0.001	6.9068
10	1.0	0.01	2.2925

[1] N. Brummer et al, "Fusion of Heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 7, Sep. 2007.

Fusion performance on SRE10 (1/3)

Actual DCF (x 1000)		I4U_1	I4U_fix01	I4U_fix02
core	itv-itv.samemic	1.312362	0.282826	0.287174
	itv-itv.diffmic	0.428066	0.675381	0.747312
	itv-tel	0.685556	0.535113	0.540624
	itv-mic	0.519078	0.493909	0.583810
	tel-tel.nve-nve	0.402542	0.440677	0.442090
	tel-tel.nve-hve	0.731301	0.789473	0.783933
	mic-mic.nvehve	0.700891	0.685236	0.835793
	tel-tel.nve-lve	0.310460	0.365771	0.342281
	mic-mic.nve-lve	0.593499	0.295308	0.461221

- Score calibration prior to fusion helps in most cases (comparing I4U_1 and I4U_fix01).
- Gender-dependent fusion (I4U_fix01) offers better performance than gender-independent fusion (I4U_fix02).

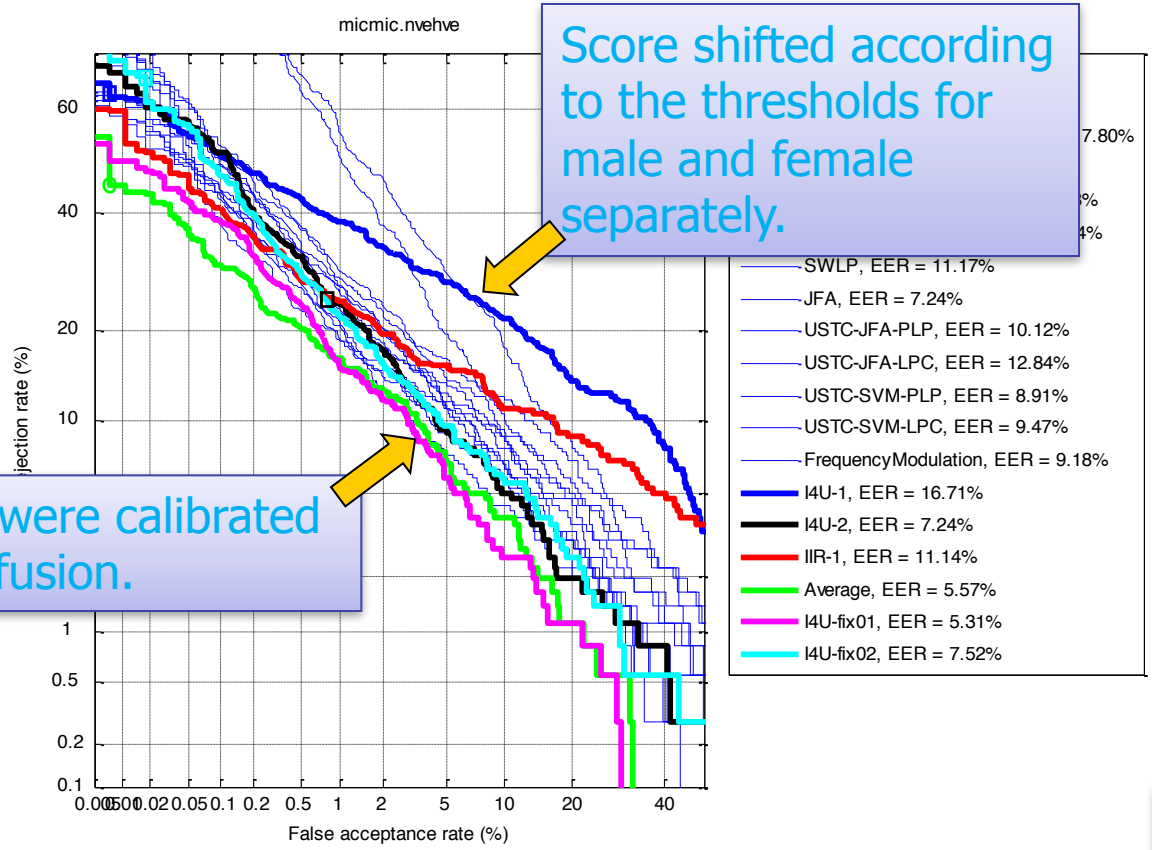
Fusion performance on SRE10 (2/3)

Actual DCF (x 10)	I4U_1	I4U_fix01	I4U_fix02
10sec-10sec	0.822689	0.798375	0.732039
core-10sec	0.359579	0.428936	0.439314
8conv-10sec	0.161062	0.191893	0.179714
8conv-core (x 1000)	0.572663	0.193429	0.184440

Fusion	Score cal	Gender Dep	Num of para
I4U_1	No	Yes	$2\times F$
I4U_fix01	Yes	Yes	$6\times F$
I4U_fix02	Yes	No	$3\times F$

- Similar results for non-core conditions.
- I4U_fix01 has more fusion (plus calibration) parameters than I4U_1 and I4U_fix02. F is the number of subsystems.

Fusion performance on SRE10 (3/3)



core-core.mic-mic.nve-hve (cond 07)				
DCF	I4U-1	I4U_fix01	I4U_fix02	Avg Weight
Actual	0.0700891	0.0685236	0.0835793	0.0802228
Min	0.0684707	0.0567716	0.0715877	0.0522618

Need further investigation here.

A*STAR I²R

Agenda

- The I4U Consortium
- System Overview
- Subsystems & Features
- Fusion Strategies & Threshold Setting
- Development Dataset Design
- Analysis of Results
- Conclusions

Conclusions

- I4U system was built upon multiple classifiers and different types of acoustic features.
- Two new SVM-based classifiers and two new features were used with acceptable performance.
- 13 sets of Dev and Eval results from this join effort (using the same set of development data) serve as a good test bed for future study of efficient score calibration and fusion methods.
- With additional score calibration, I4U submission could be improved in some tasks.