Classifier Subset Selection and Fusion for Speaker Verification

Filip Sedlák[1], Tomi Kinnunen[1], Ville Hautamäki[2], Kong-Aik Lee[2], Haizhou Li[1,2]

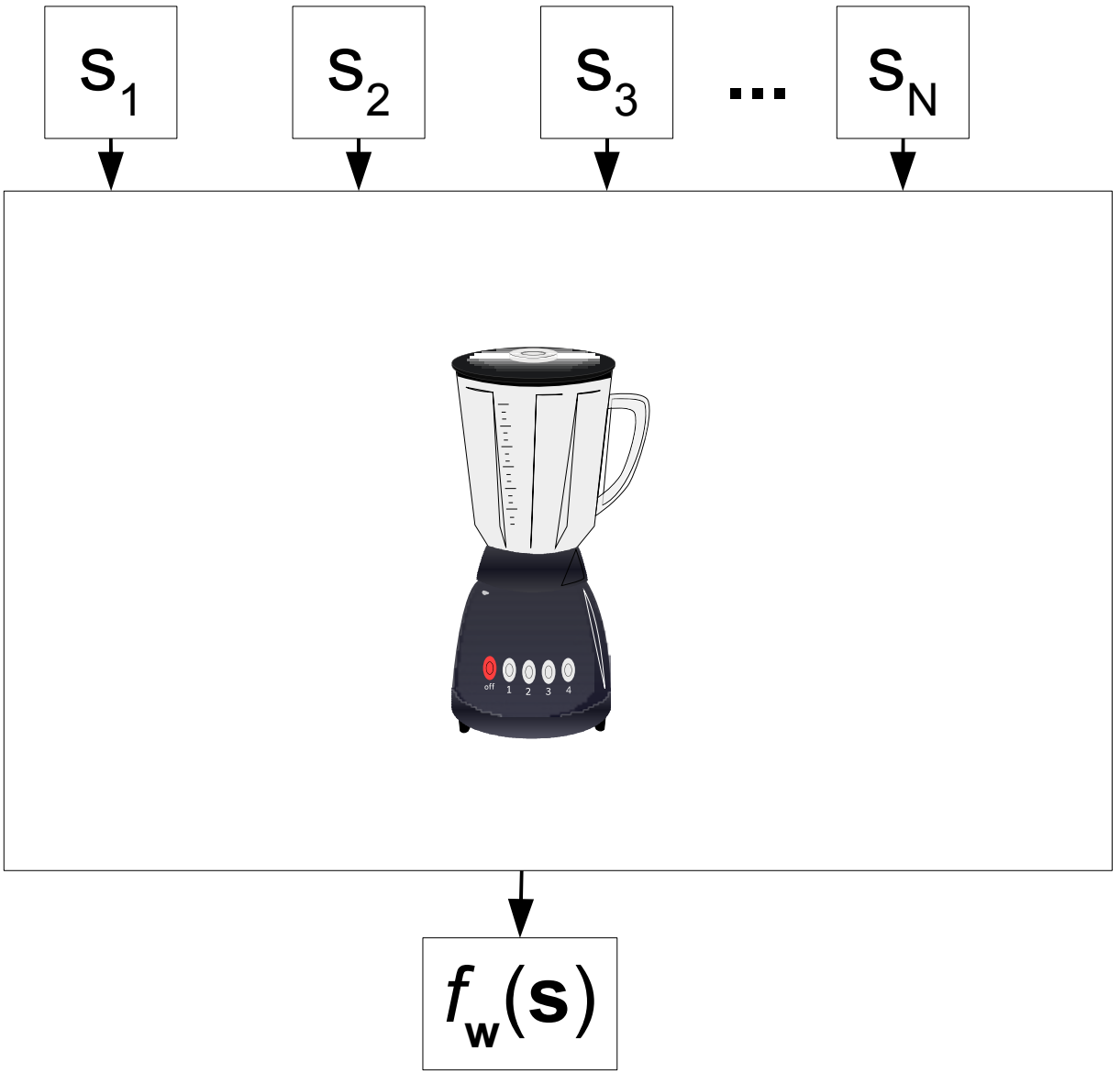
[1]University of Eastern Finland, School of Computing, Joensuu, Finland

[2]Human Language Technology Department, Institute of Infocomm Research (I²R), A*STAR, Singapore

Classifier Fusion

Verdict?

Classifier Fusion

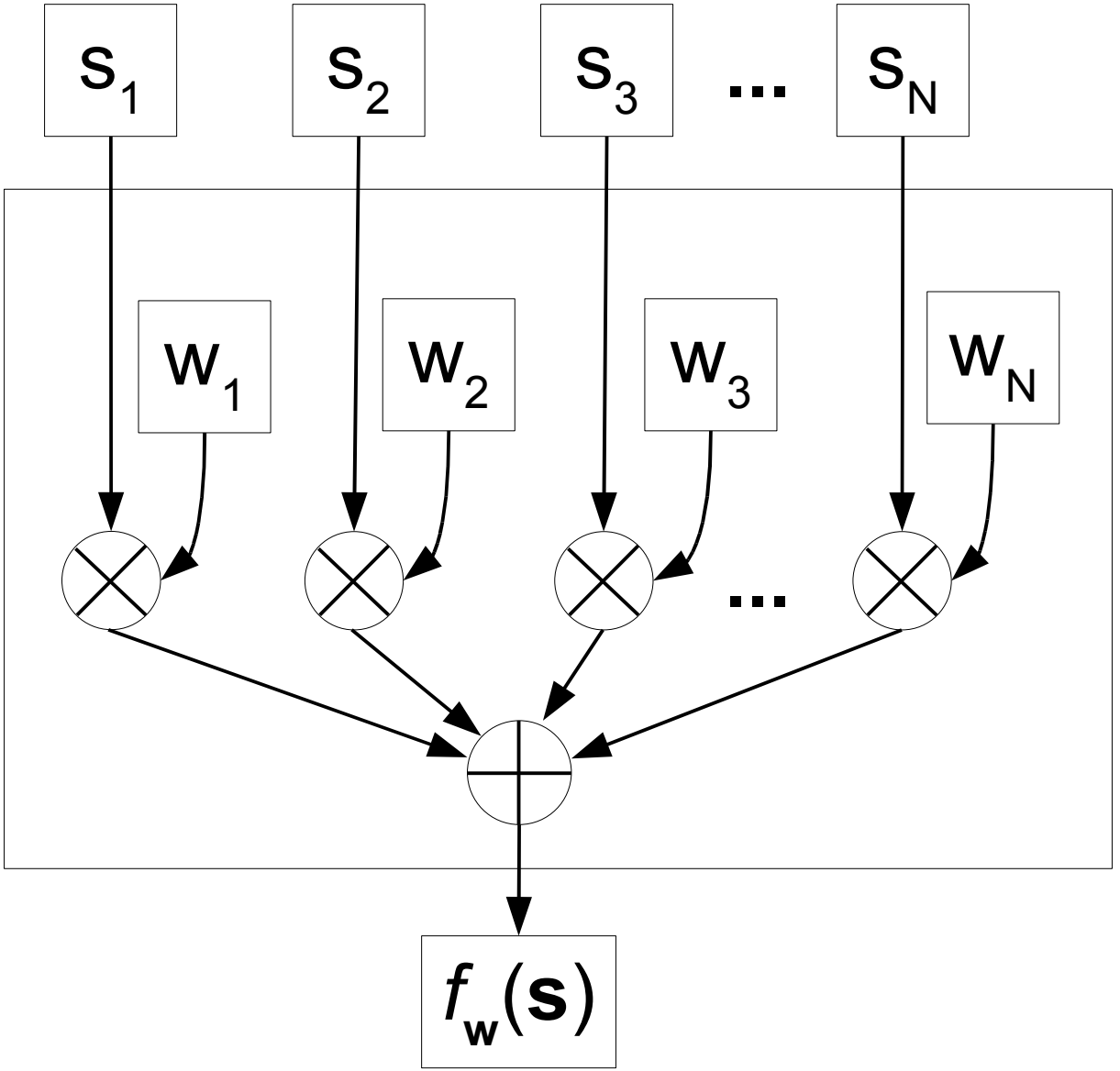


Scores from N classifiers

Fusion

Fused score

Linear Fusion



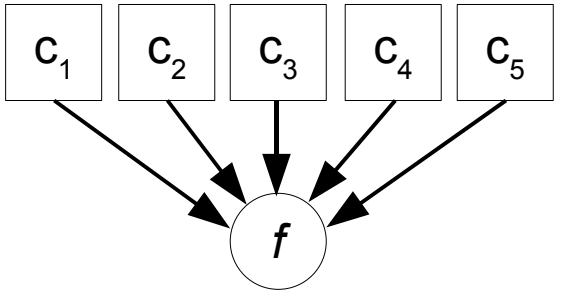
Scores from N classifiers

Trained fusion weights

$$f_w(\boldsymbol{s}) = \sum_{k=1}^{N} w_k s_k$$

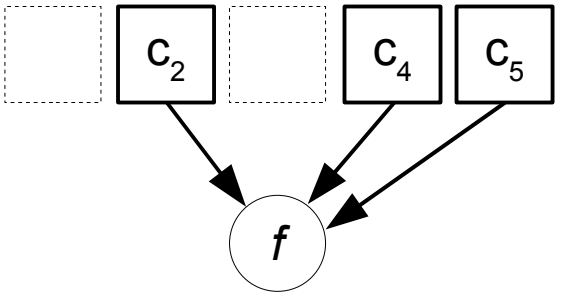
Full Set vs. Subset

Full set of N=5 classifiers

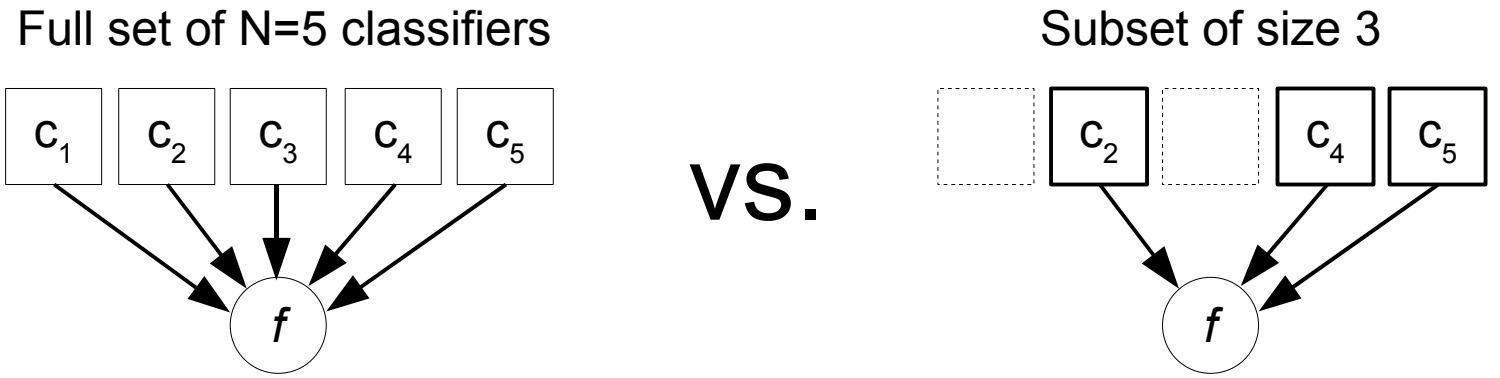


vs.

Subset of size 3



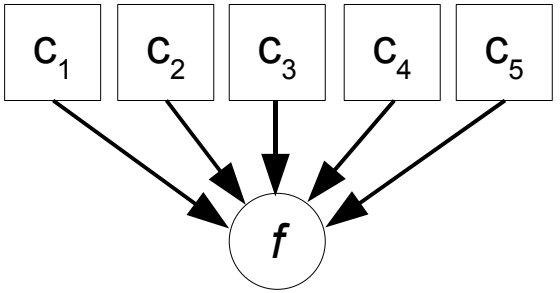
Full Set vs. Subset



+ straightforward
+ computationally efficient
- possibly over-fitting if N is large

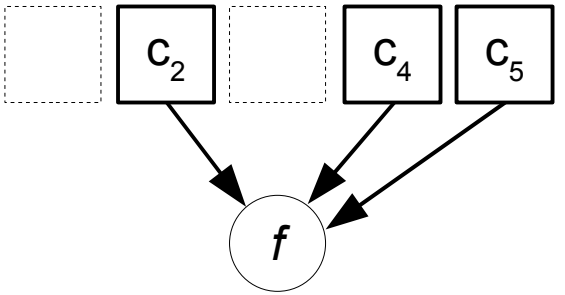
Full Set vs. Subset

Full set of N=5 classifiers



vs.

Subset of size 3

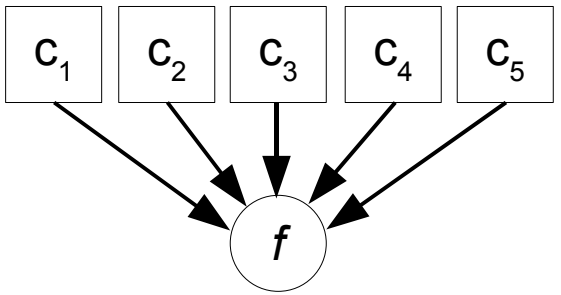


+ straightforward
+ computationally efficient
- possibly over-fitting if N is large

+ possibly better generalization
- requires subset selection

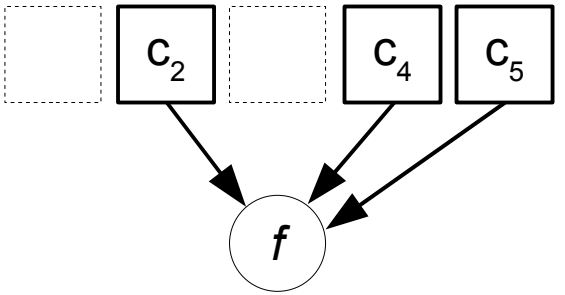
Full Set vs. Subset

Full set of N=5 classifiers



vs.

Subset of size 3

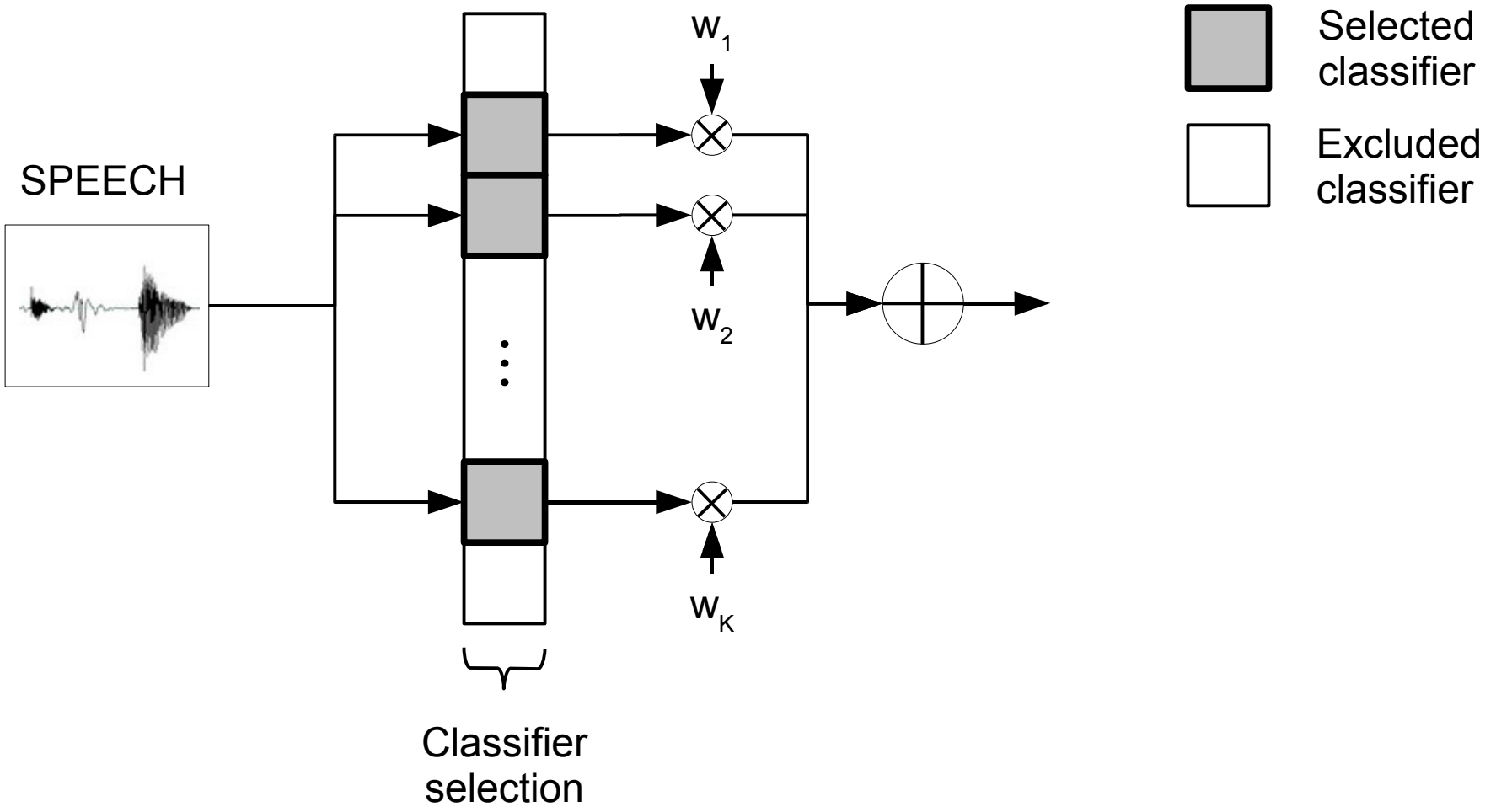


- \+ straightforward
- \+ computationally efficient
- \- possibly over-fitting if N is large

- \+ possibly better generalization
- \- requires subset selection

Can a subset fusion give better performance than the full-set?

Joint Classifier Selection and Fusion



Joint Classifier Selection and Fusion

Train **S-Cal** [1] score warping for each individual subsystem

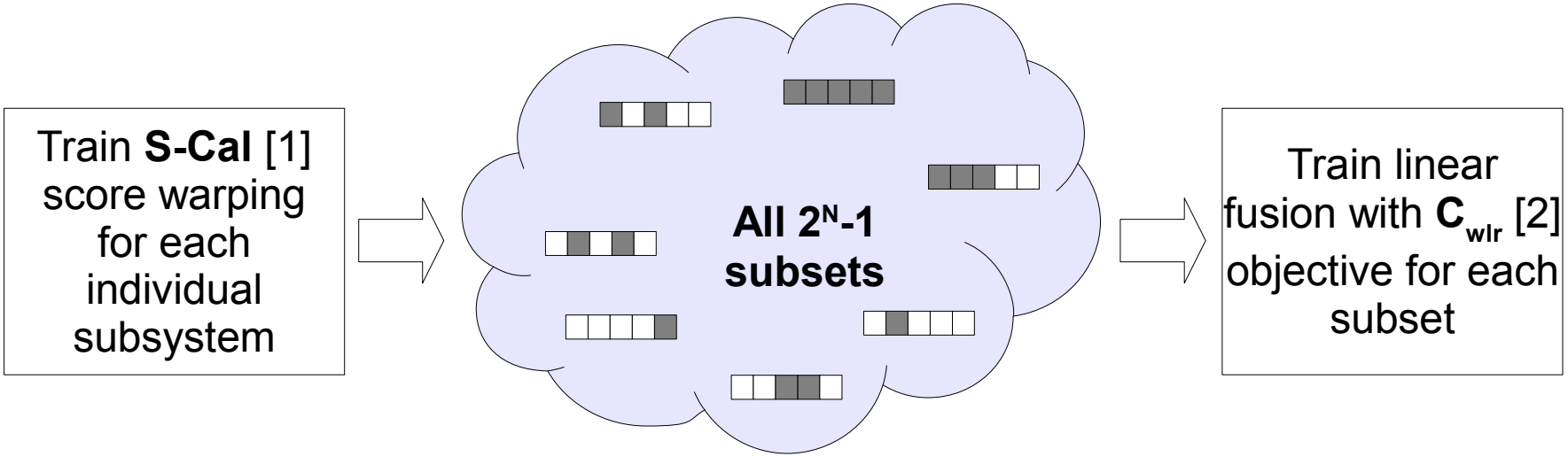
$$\mathrm{SCal}(s) = \log \frac{(\mathrm{logit}^{-1}\alpha)(\mathrm{e}^{as+b}-1)+1}{(\mathrm{logit}^{-1}\beta)(\mathrm{e}^{as+b}-1)+1}$$

$$\mathrm{logit}(x) = \log \frac{x}{1-x}$$

$$C_{\mathrm{llr}} = \frac{1}{K}\sum_{i=1}^{K}\log\left(1+\mathrm{e}^{-s_{Target}}\right) + \frac{1}{L}\sum_{j=1}^{L}\log\left(1+\mathrm{e}^{s_{Non-Target}}\right)$$

[1] Niko Brummer and Johan du Preez, “Application-Independent Evaluation of Speaker Detection”, Computer Speech and Language, 2005.

Joint Classifier Selection and Fusion



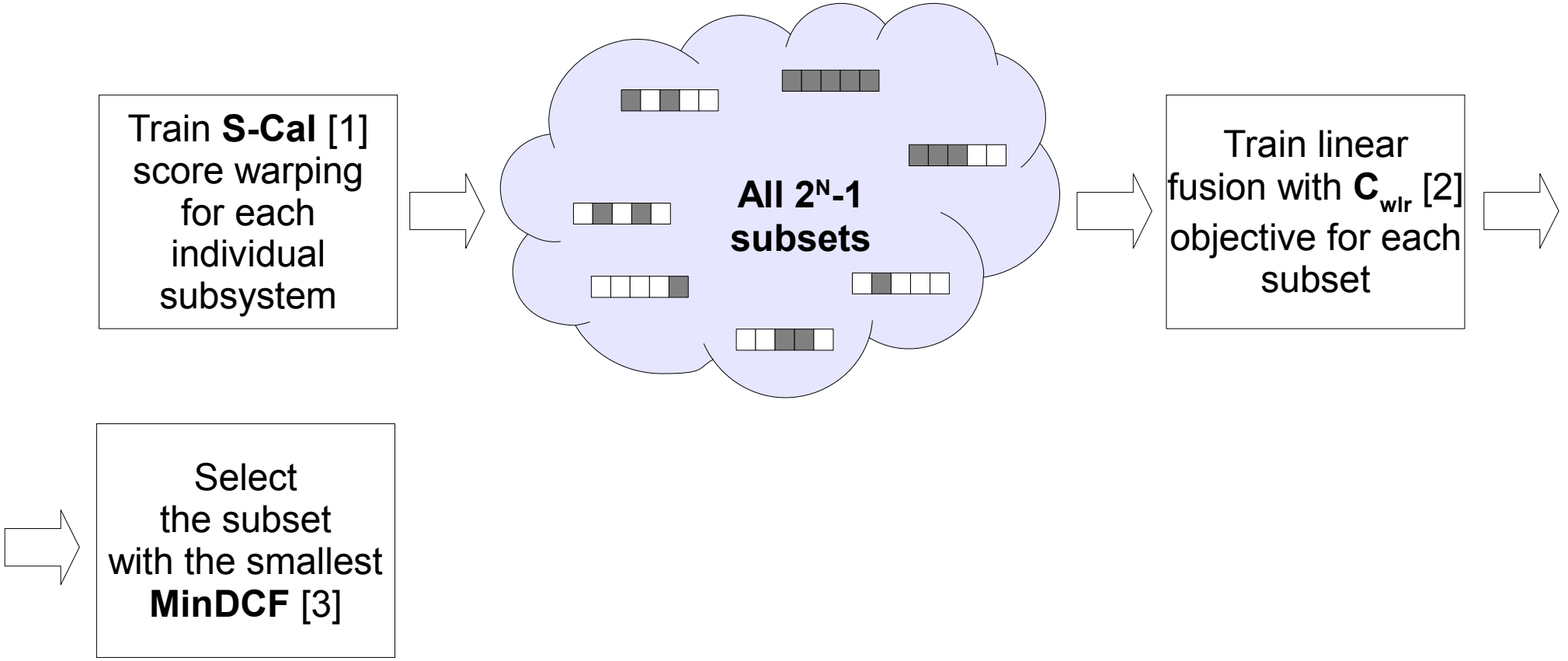
$$C_{\text{wlr}} = \frac{P}{K}\sum_{i=1}^{K}\log\left(1+\mathrm{e}^{-s_{Target}-\text{logit}(P)}\right)+\frac{1-P}{L}\sum_{j=1}^{L}\log\left(1+\mathrm{e}^{s_{Non-Target}+\text{logit}(P)}\right)$$

$$P = \text{logit}^{-1}\left(\text{logit}(P_{\text{Target}})+\log\frac{C_{\text{Miss}}}{C_{\text{FA}}}\right)$$

Here $C_{Miss} = 1$, $C_{FA} = 1$, $P_{Target} = 0.001$ (i.e. the 'new' NIST cost function).

[2] Niko Brummer, "Application-Independent Evaluation of Speaker Detection", Odyssey 2004.

Joint Classifier Selection and Fusion



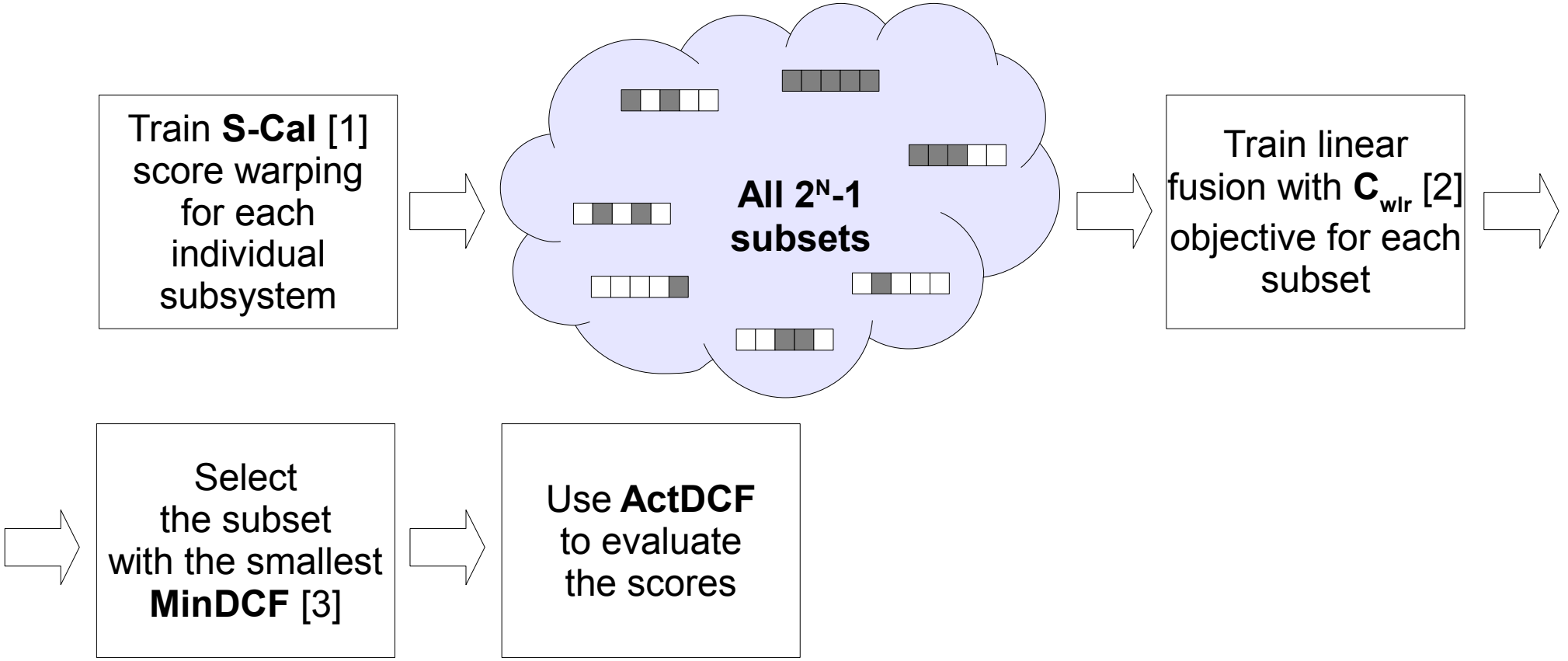
$$\mathrm{DCF}(\theta) = C_{\mathrm{Miss}} P_{\mathrm{Miss}}(\theta) P_{\mathrm{Target}} + C_{\mathrm{FA}} P_{\mathrm{FA}}(\theta)(1 - P_{\mathrm{Target}})$$

$$\mathrm{MinDCF} = \min_{\theta} \mathrm{DCF}(\theta)$$

Here $C_{Miss} = 1$, $C_{FA} = 1$, $P_{Target} = 0.001$ (i.e. the 'new' NIST cost function).

[3] http://www.itl.nist.gov/iad/mig/tests/sre/

Joint Classifier Selection and Fusion



$$\mathrm{DCF}(\theta) = C_{\mathrm{Miss}} P_{\mathrm{Miss}}(\theta) P_{\mathrm{Target}} + C_{\mathrm{FA}} P_{\mathrm{FA}}(\theta)(1 - P_{\mathrm{Target}})$$

$$\mathrm{ActDCF} = \mathrm{DCF}(\theta_{\mathrm{Trainset}})$$

Here C_{Miss} = 1, C_{FA} = 1, P_{Target} = 0.001 (i.e. the 'new' NIST cost function).

[3] http://www.itl.nist.gov/iad/mig/tests/sre/

Base Classifiers (I4U, NIST 2010)

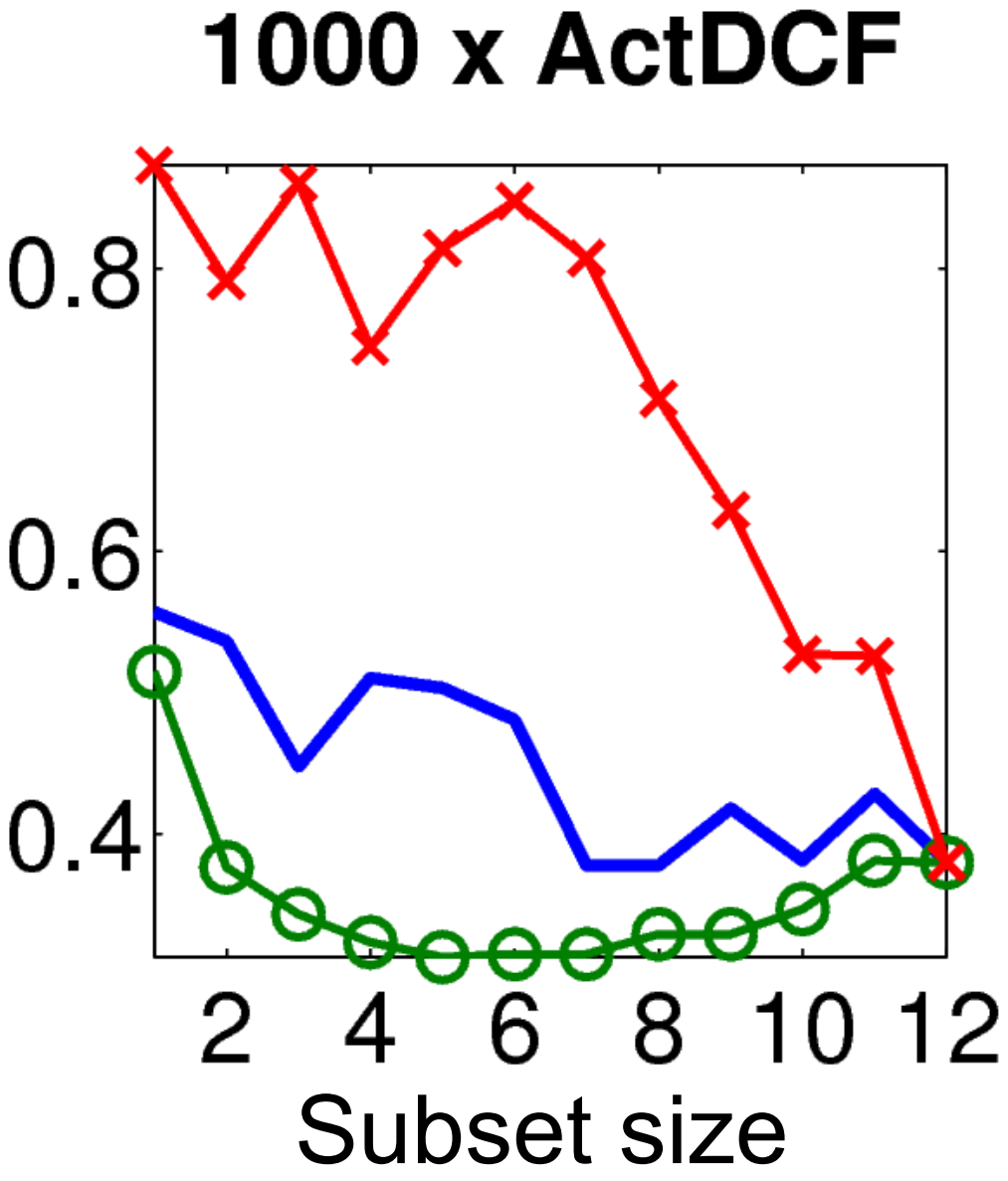
	Classifier	Feature	Evalset1 EER (%)	Evalset2 EER (%)
1	GMM-UBM	PLP	**3.95**	4.99
2	Joint Factor Analysis	PLP	4.24	4.12
3		PLP	4.24	3.75
4		LPCC	4.59	5.74
5	GMM-SVM	PLP	5.65	5.49
6	Kullback–Leibler Divergence	MFCC	4.99	4.37
7		LPCC	6.45	5.37
8		MLF	5.81	4.74
9		LPCC	4.24	6.52
10		SWLP	10.20	5.87
11	GMM-SVM Feature Transformation	PLP	8.13	6.12
12	GMM-SVM Bhattacharyya Distance	PLP	5.40	**3.03**

All datasets:
interview-telephone, female trials

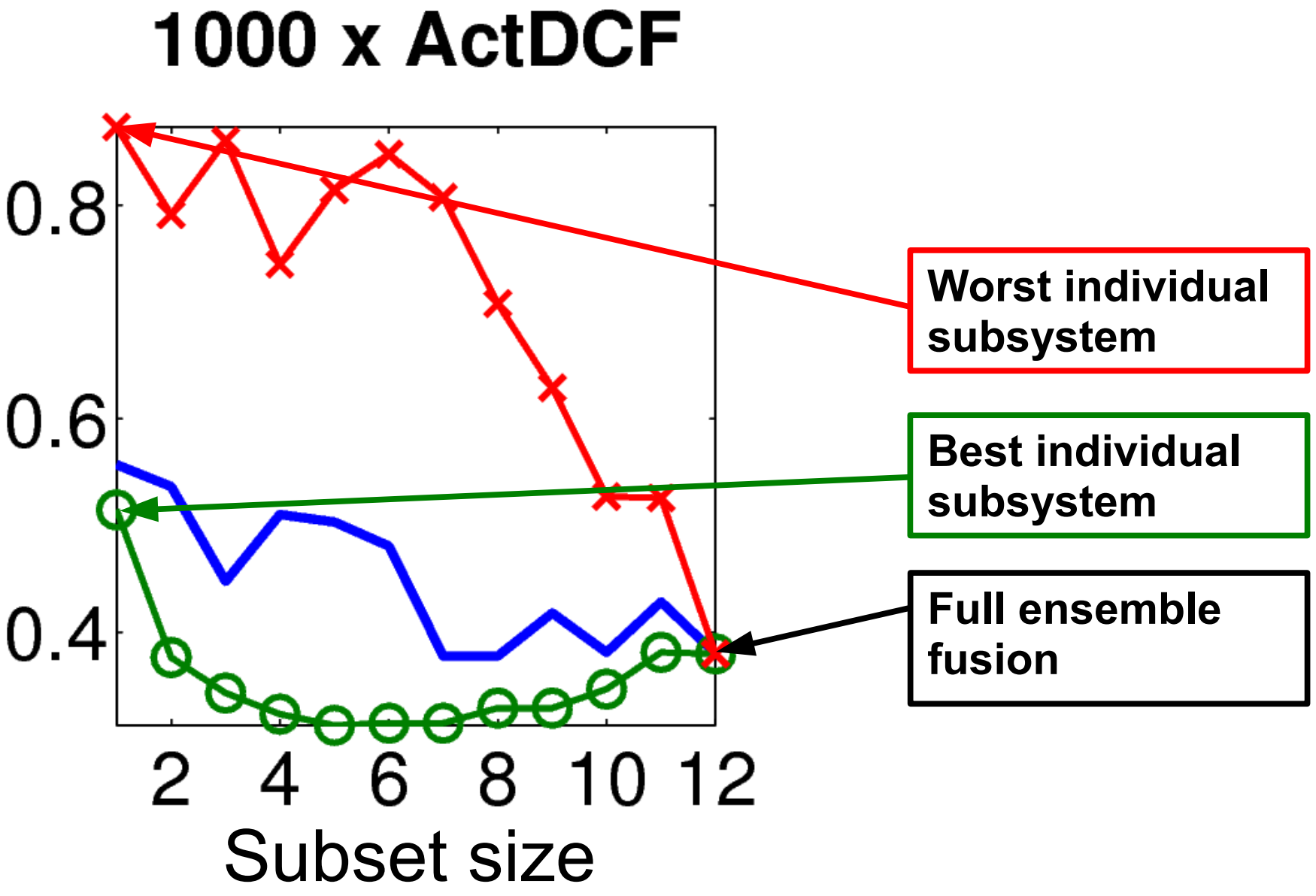
Trainset, Evalset1:
NIST SRE 2008

Evalset2:
NIST SRE 2010

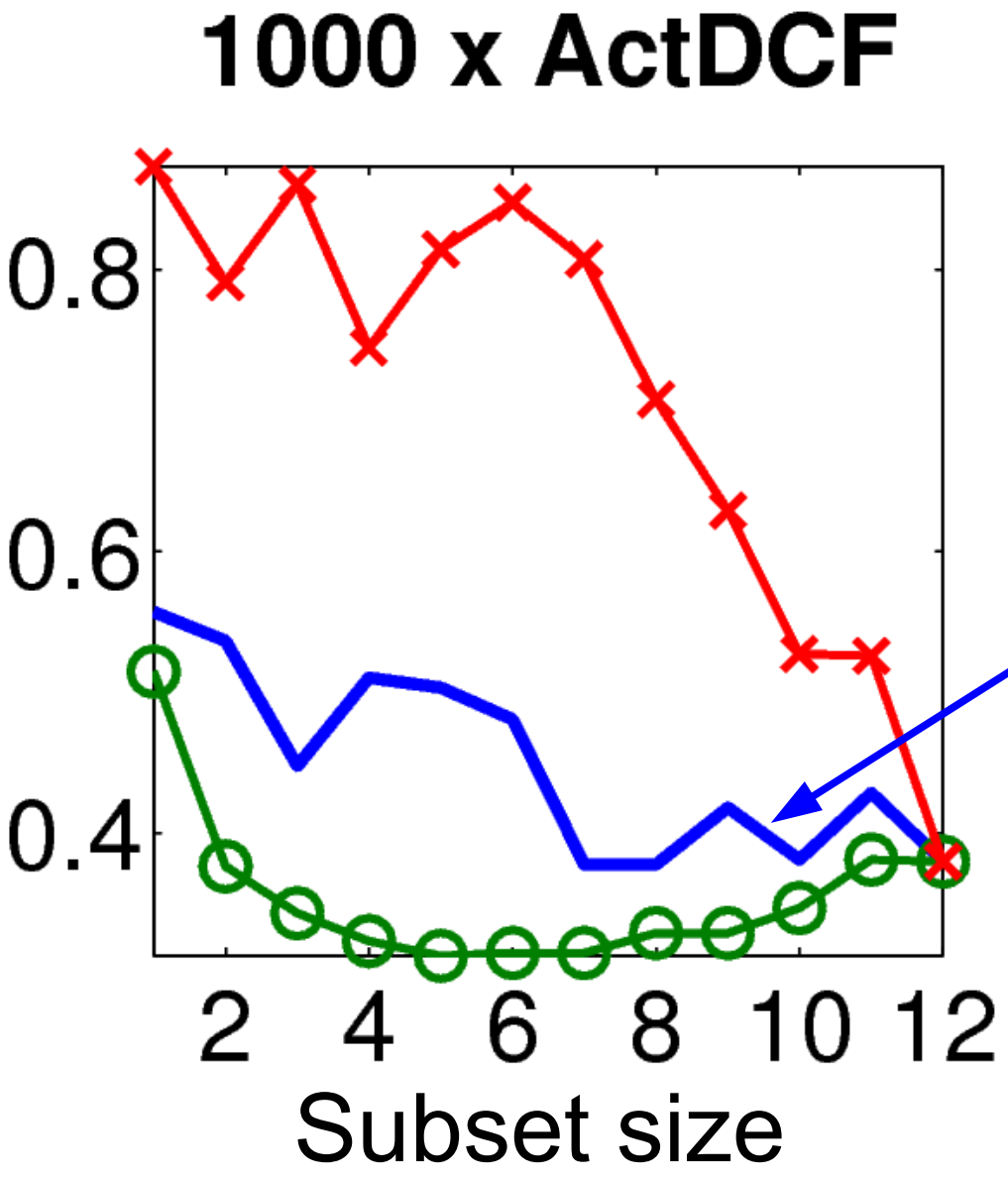
Error Bounds on the NIST2008 (Evalset1)



Error Bounds on the NIST2008 (Evalset1)

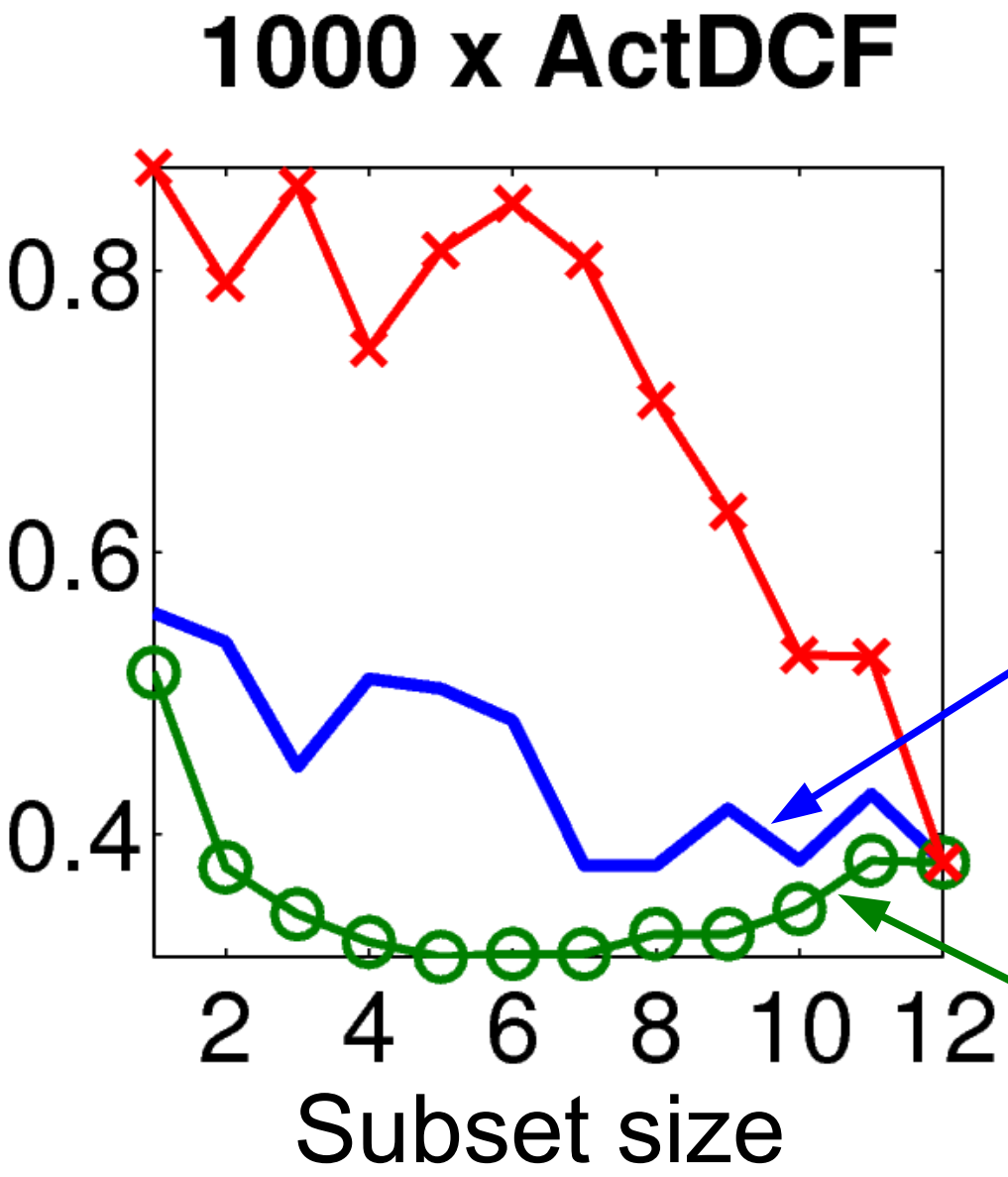


Error Bounds on the NIST2008 (Evalset1)



Realistic use-case
Fusion is trained on the **trainset**, subset is selected based on the **best** performance on the **trainset**.

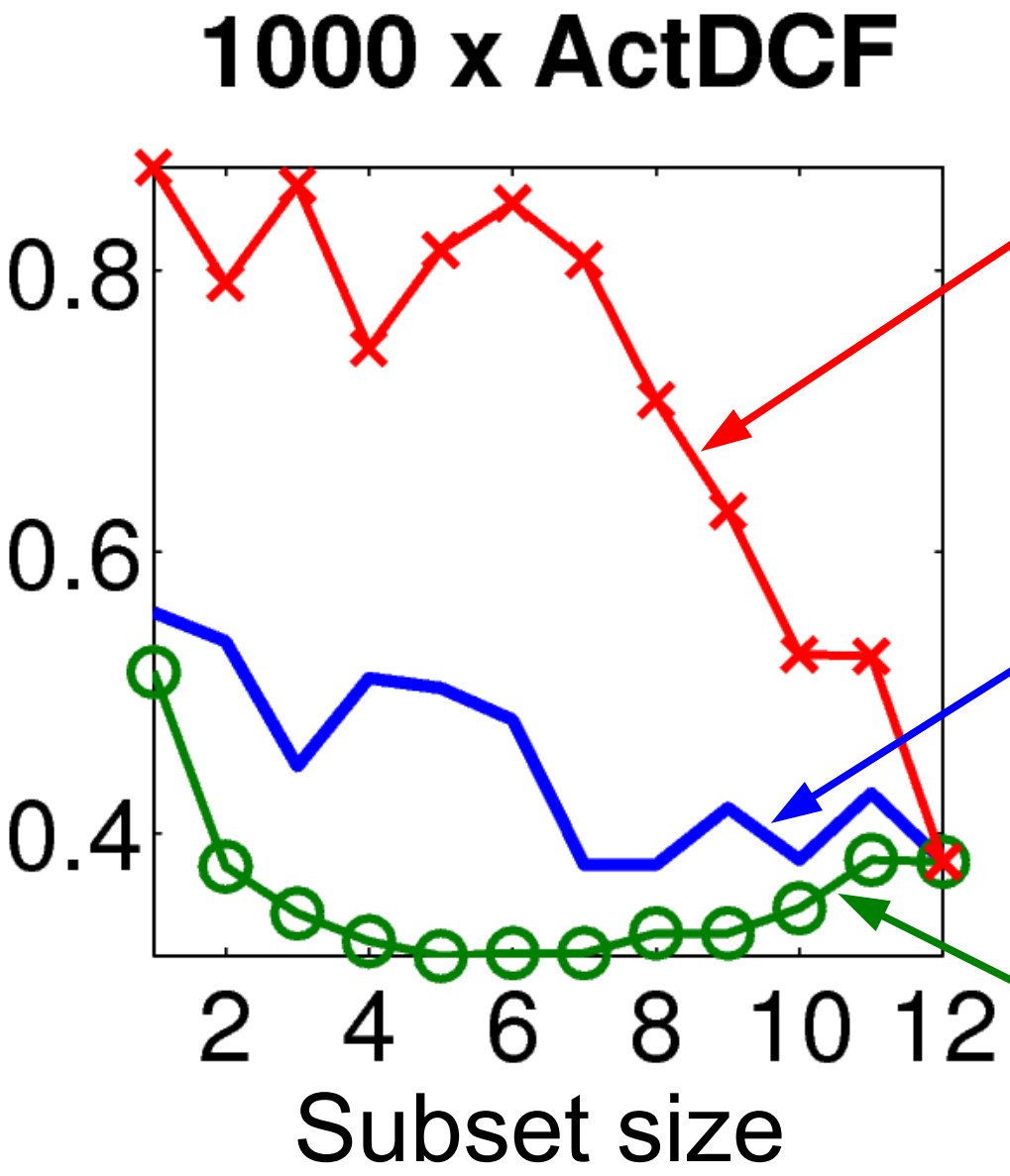
Error Bounds on the NIST2008 (Evalset1)



Realistic use-case
Fusion is trained on the **trainset**, subset is selected based on the **best** performance on the **trainset**.

Best subset selection oracle
Fusion is trained on the **trainset**, subset is selected based on the **best** performance on the **evalset1**.

Error Bounds on the NIST2008 (Evalset1)

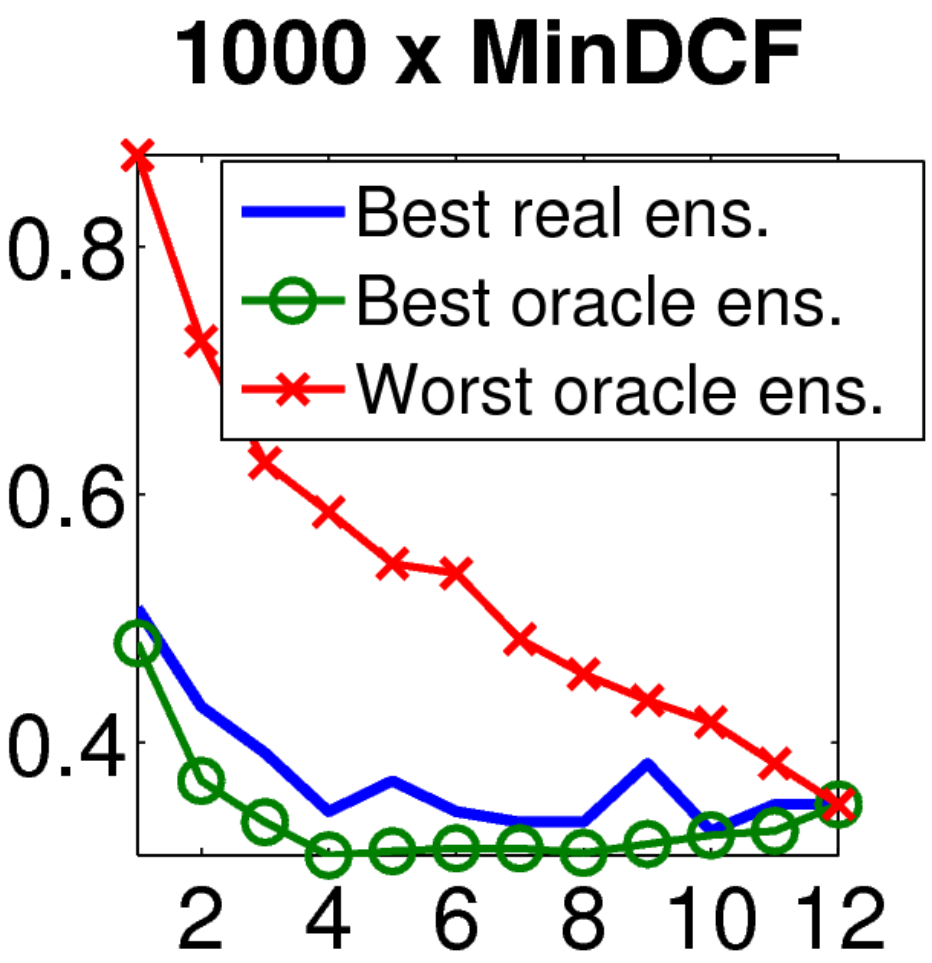


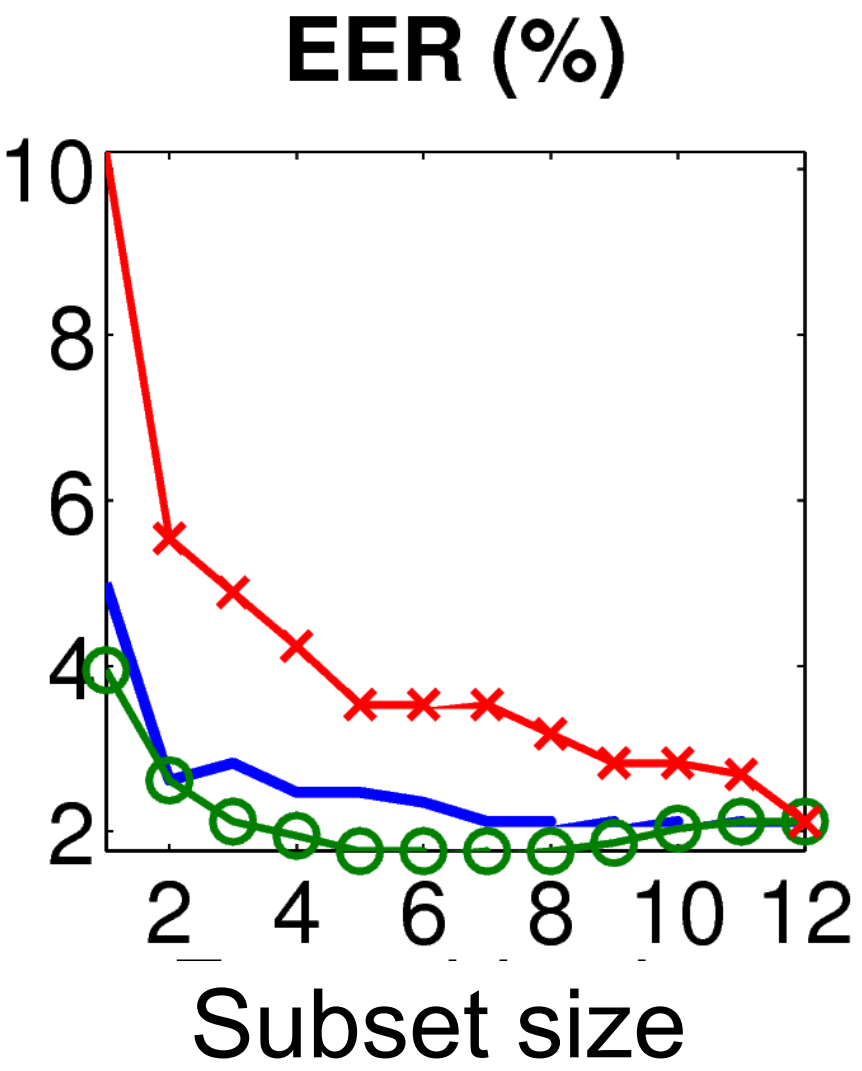
Worst subset selection oracle
Fusion is trained on the **trainset**, subset is selected based on the **worst** performance on the **evalset1**.

Realistic use-case
Fusion is trained on the **trainset**, subset is selected based on the **best** performance on the **trainset**.

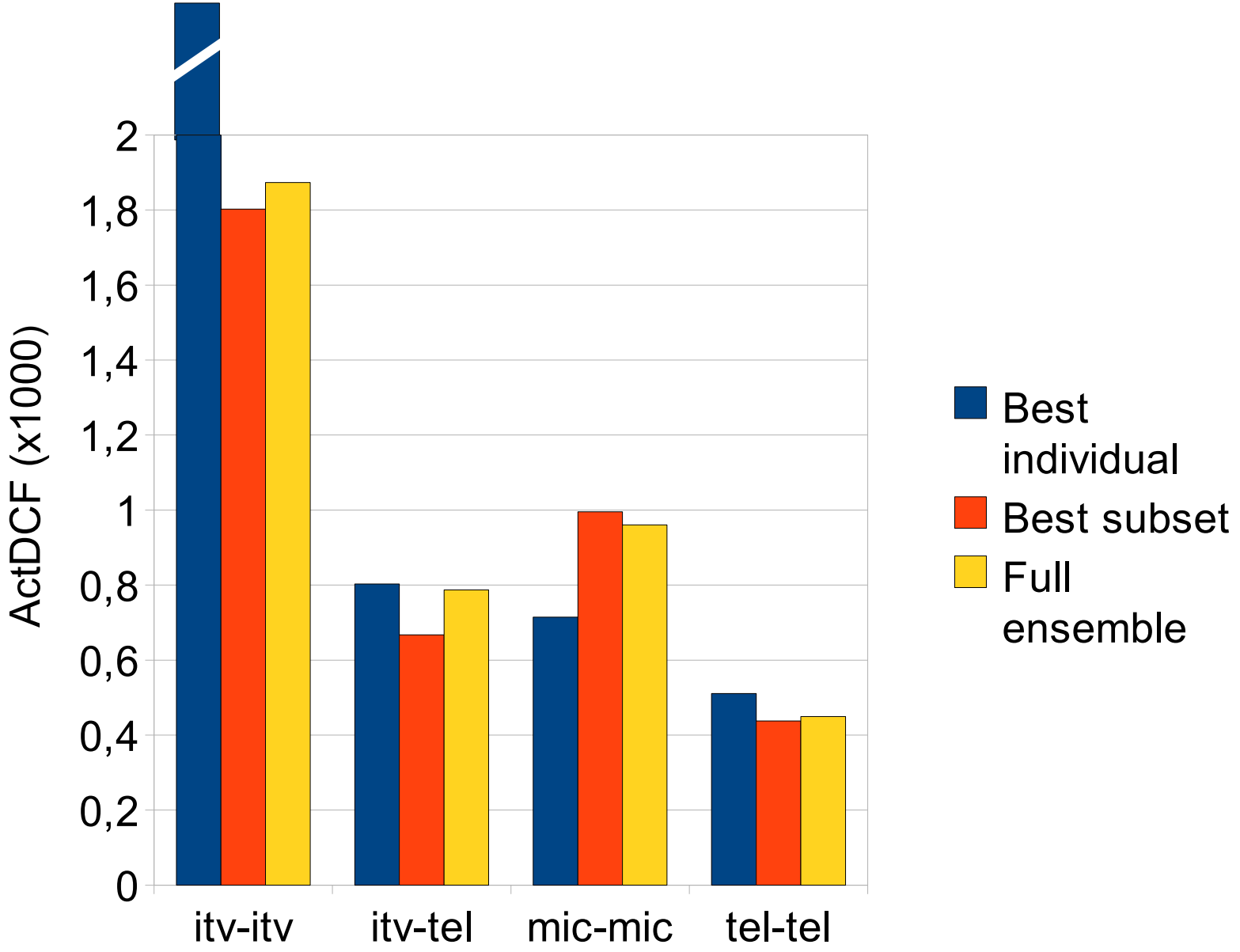
Best subset selection oracle
Fusion is trained on the **trainset**, subset is selected based on the **best** performance on the **evalset1**.

Error Bounds on the NIST2008 (Evalset1)



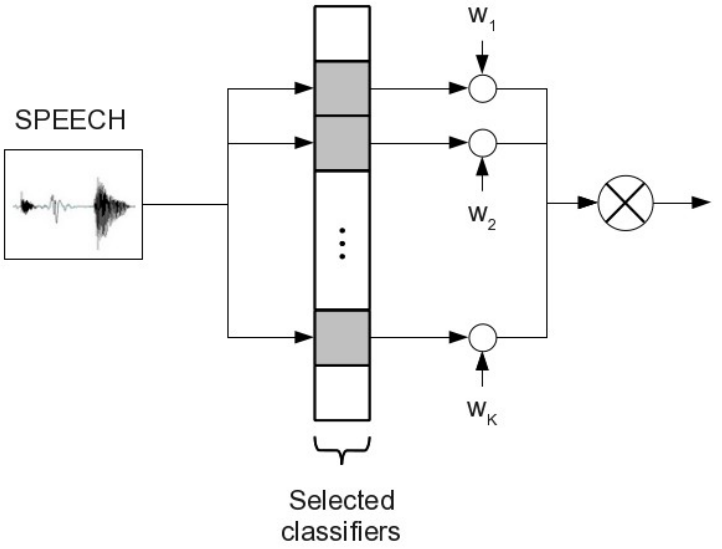


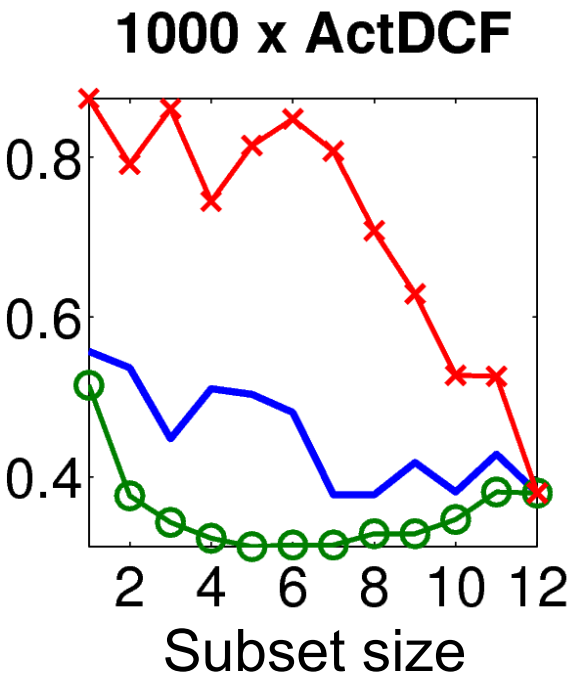
Subset Performance on Evalset2 (NIST2010), Pooled Genders



Conclusions

- Subset fusion has the potential to outperform the full set fusion.
- Further study should focus on subset selection methods.





Following are support slides...

Score Sets

	Trainset	Devset	Evalset
Trials	NIST2008 itv-tel female	NIST2008 itv-tel female	NIST2010 itv-tel female
Target	263	283	801
Non-target	27 315	27 195	30 254

Error Bounds

- Best individual base system
- Full set fusion

Weights trained on	Best subset selected on	Performance evaluated on	
Trainset	Trainset	Devset/Evalset	'Real'
Trainset	Devset/Evalset	Devset/Evalset	'Best Real'
Devset/Evalset	Devset/Evalset	Devset/Evalset	'Best Oracle'