Fusion of Several Binary Classifiers for Countermeasure of Speech Replay Spoofing Attack

Andrey A. Lependin, Yacob A. Filin

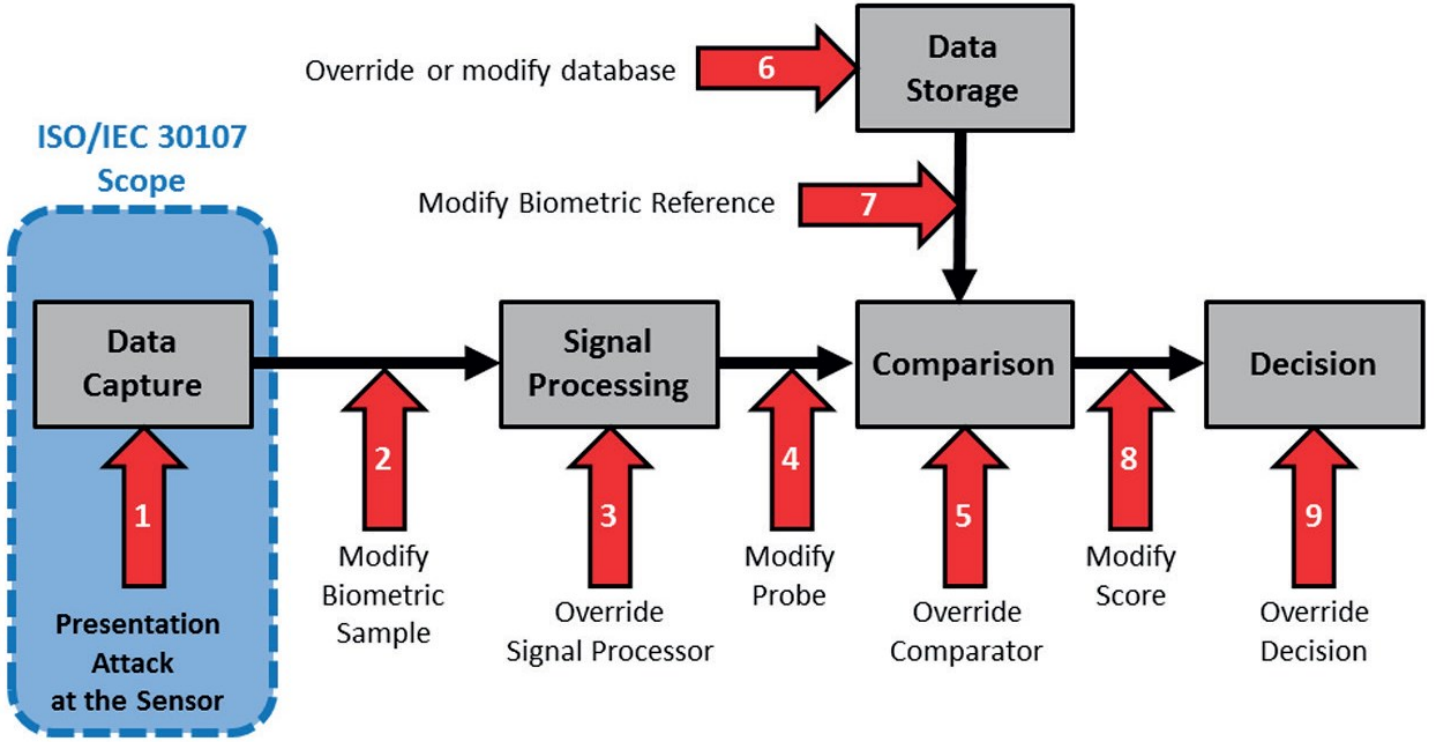
Altai State University, Barnaul, Russia

Overview

- Introduction
- Features and post-processing
- Classifiers
- Dataset
- Performance evaluation
- Results
- Conclusion

Introduction. Spoofing

Presentation attacks [ISO/IEC 30107-1:2016]



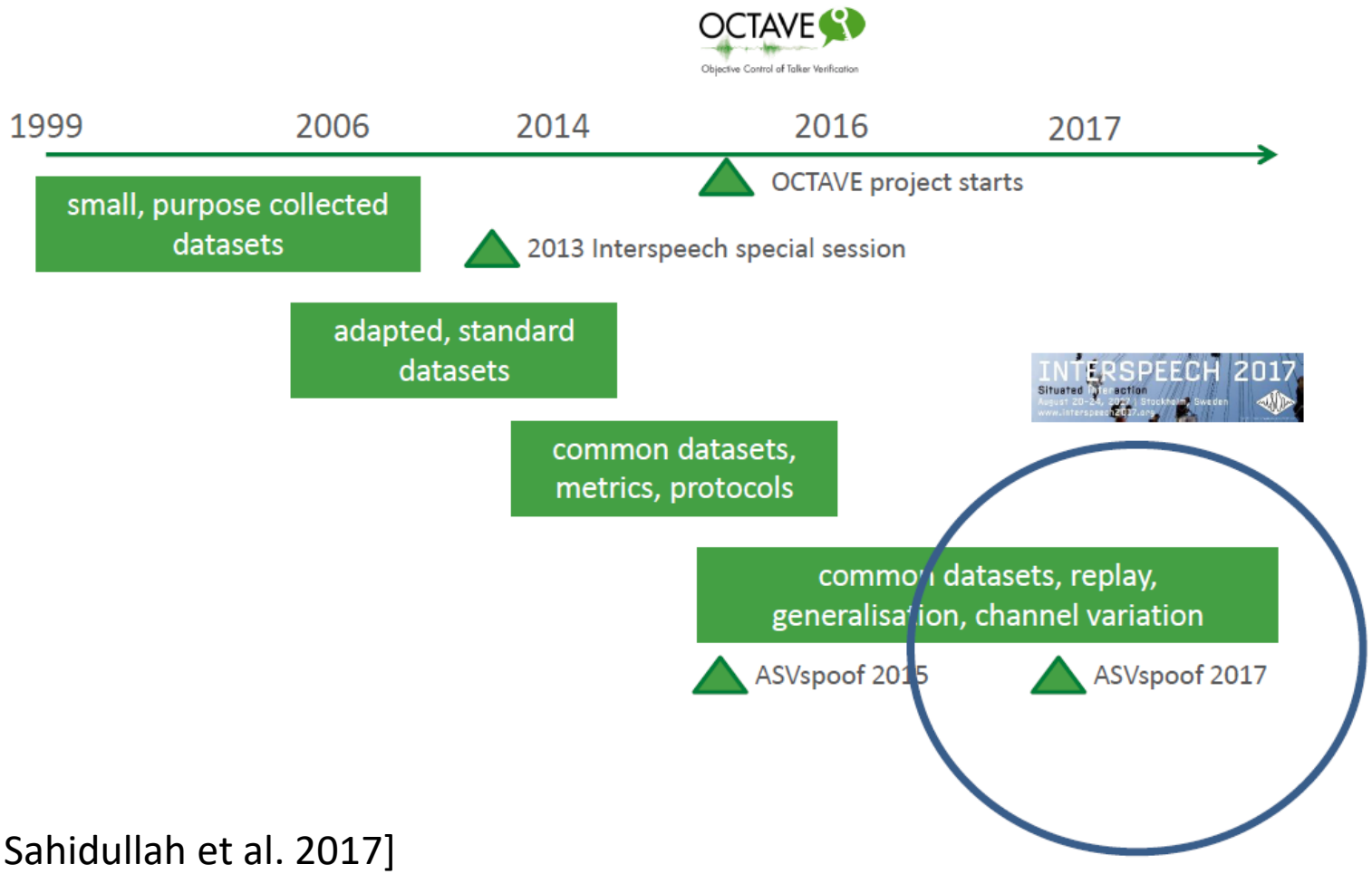
Replay Attacks

1. **Phrase prompting with utterance verification**
   Did the user speak the prompted text ?
2. **Audio fingerprinting**
   Do I know this recording ?
3. **Speaker-independent replay detection**
   Is this recording authentic or replayed one ?

ASVspoof 2017

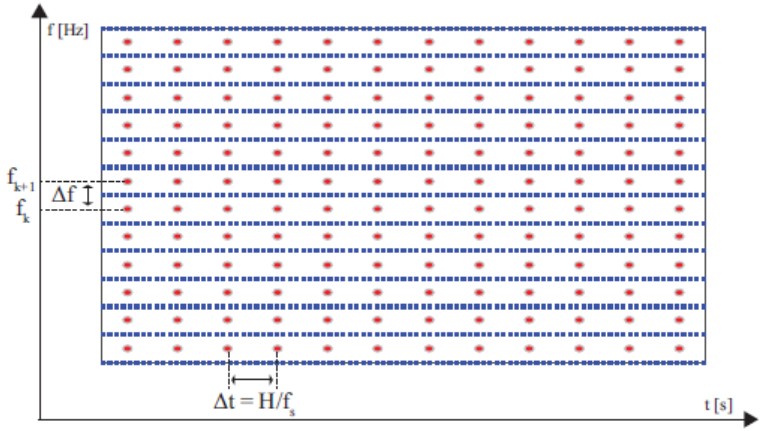
[T. Kinnunen, M. Sahidullah et al. 2017]

ASV Spoof 2017 Challenge

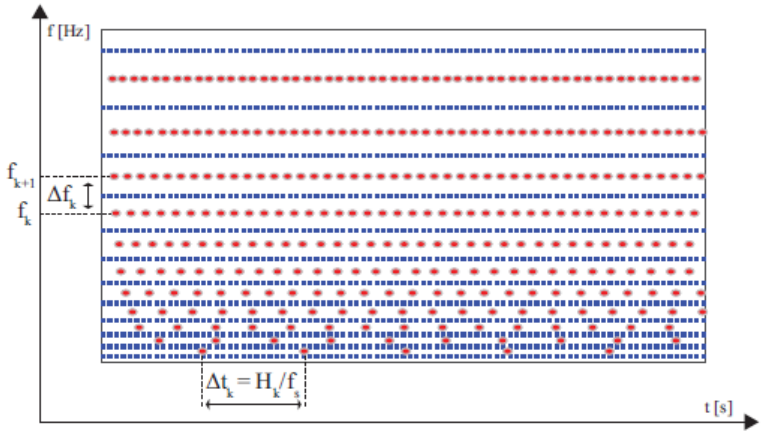


[T. Kinnunen, M. Sahidullah et al. 2017]

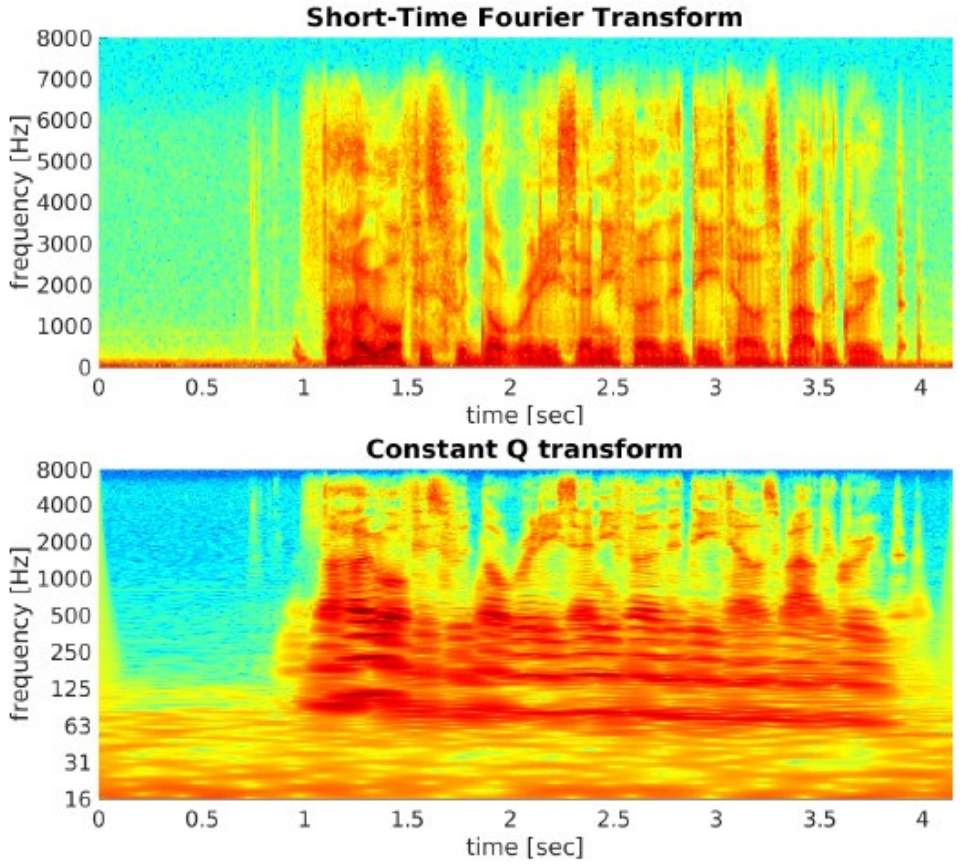
Constant-Q Transform (CQT)



(a) STFT

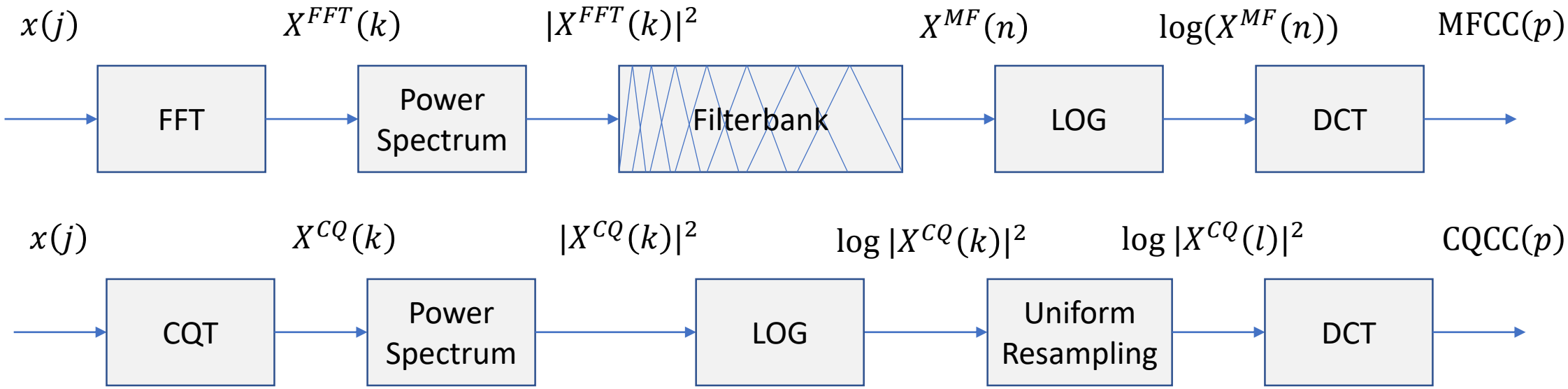


(b) CQT



[M. Todisco, H. Delgado, N. Evans. 2016]

MFCC vs CQCC



$$X^{CQ}(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j)a_k^*(j-n+N_k/2)$$

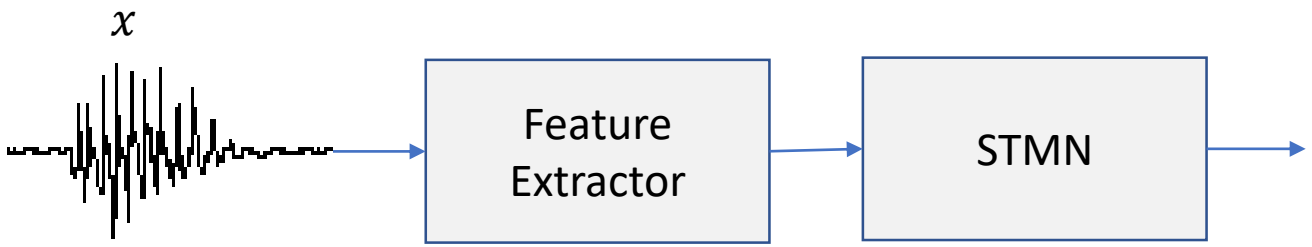
$$a_k(n) = \frac{1}{C}\left(\frac{n}{N_k}\right)exp\left[i\left(2\pi n\frac{f_k}{f_s} + \Phi_k\right)\right]$$

where $f_k = f_1 2^{\frac{k-1}{B}}$

$$CQCC(p) = \sum_{l=1}^{L} log|X^{CQ}(l)|^2 cos\left[\frac{p(l-\frac{1}{2})\pi}{L}\right]$$

[M. Todisco, H. Delgado, N. Evans. 2016]

Feature Normalization (STMN)



Short-time Mean and Variance Normalization (STMVN)

$$C_{stmvn}(m,k)=\frac{C(m,k)-\mu_{st}(m,k)}{\sigma_{st}(m,k)},$$

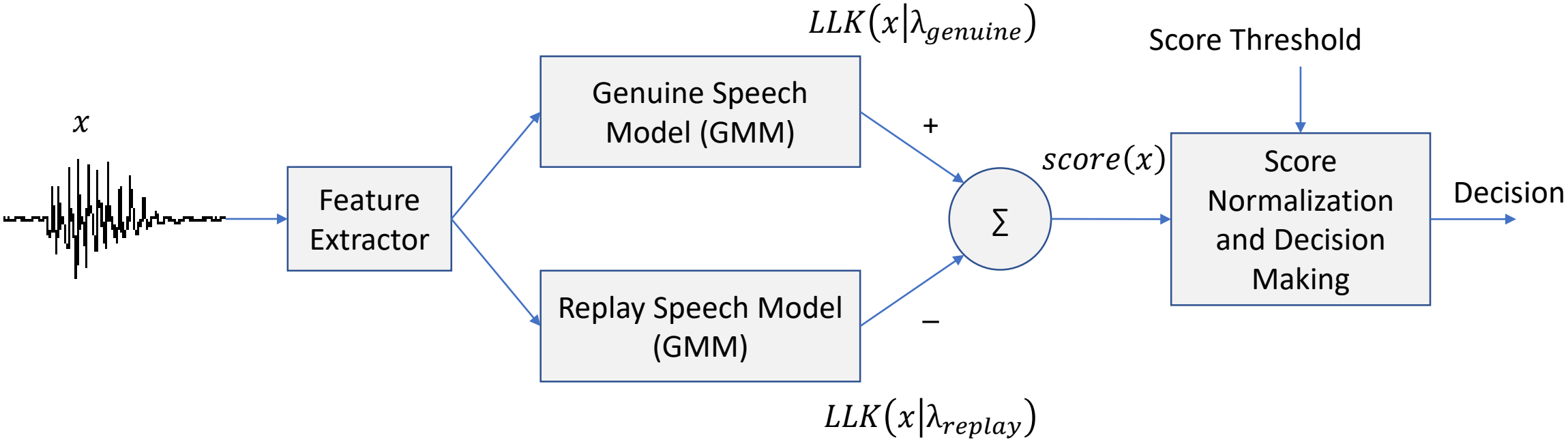
$$\mu_{st}(m,k)=\frac{1}{L}\sum_{j=m-L/2}^{m+L/2}C(j,k)$$

$$\sigma_{st}(m,k)=\frac{1}{L}\sum_{j=m-L/2}^{m+L/2}\left(C(j,k)-\mu(m,k)\right)^2$$

[J. Alam, P. Ouellet et al. 2011]

$$C_{STMN}(m,k) = C(m,k)- \mu_{ST}(m,k),$$
$$\sigma_{ST}(m,k) = 1$$

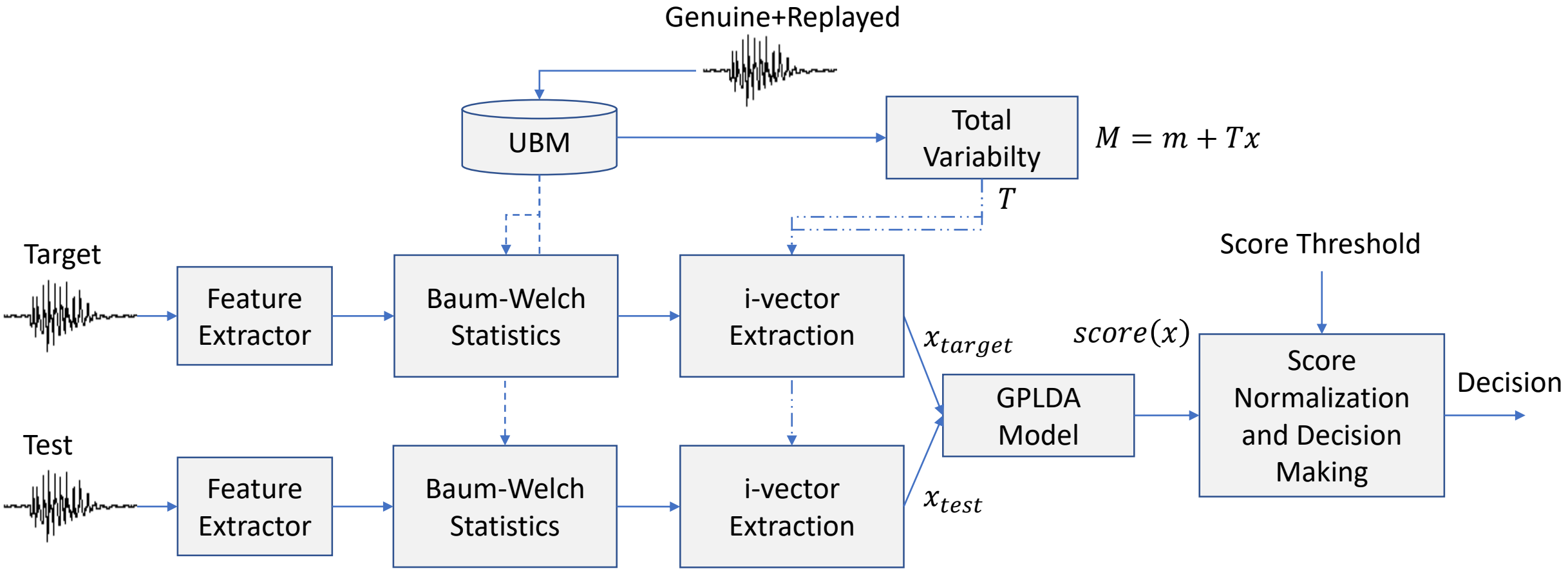
ASVSpoof2017 Baseline Classifier



$$score(x) = LLK(x|\lambda_{genuine}) - LLK(x|\lambda_{replay})$$

$$LLK(x|\lambda) = \frac{1}{T}\sum_{i=1}^{T} \log p(x_i|\lambda)$$

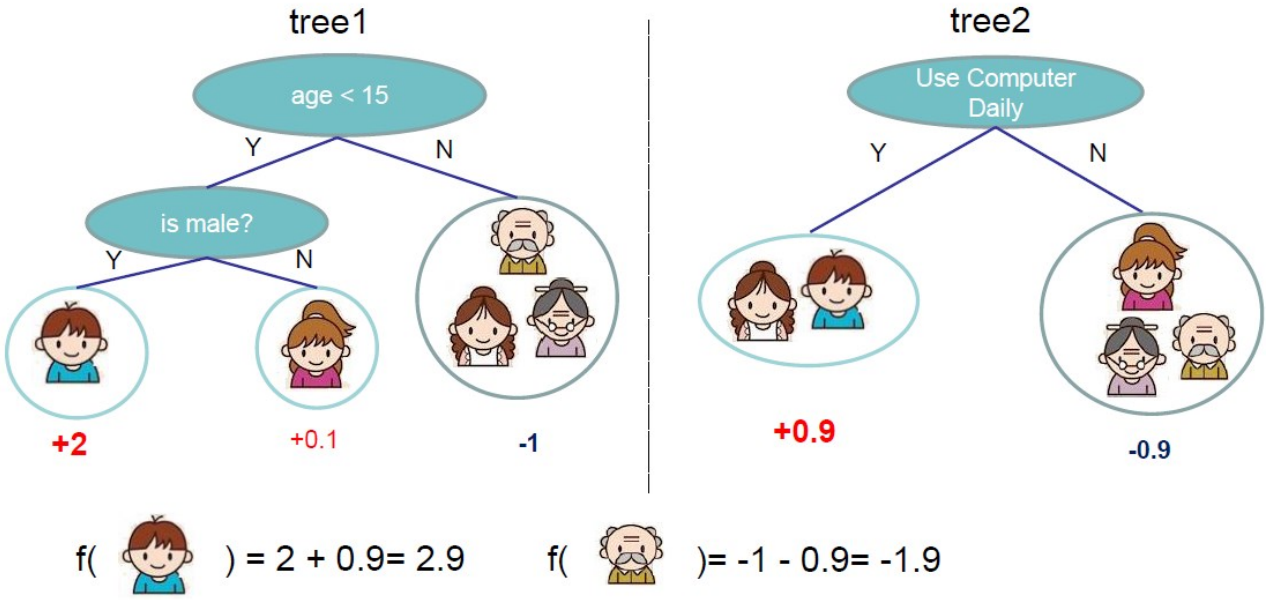
i-vector extraction and GPLDA



$$LLR(x_{target}, x_{test}) = \log \frac{P(x_{target}, x_{test}|H_1)}{P(x_{target}|H_0)P(x_{test}|H_0)}$$

$$score(x) = \frac{1}{\#G}\sum_{x_g \in G} LLR(x_g, x) - \frac{1}{\#R}\sum_{x_r \in R} LLR(x_r, x)$$

Gradient Boosting (XGBoost)



- Model: assuming we have K trees

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

- Objective

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Training loss — Complexity of the Trees

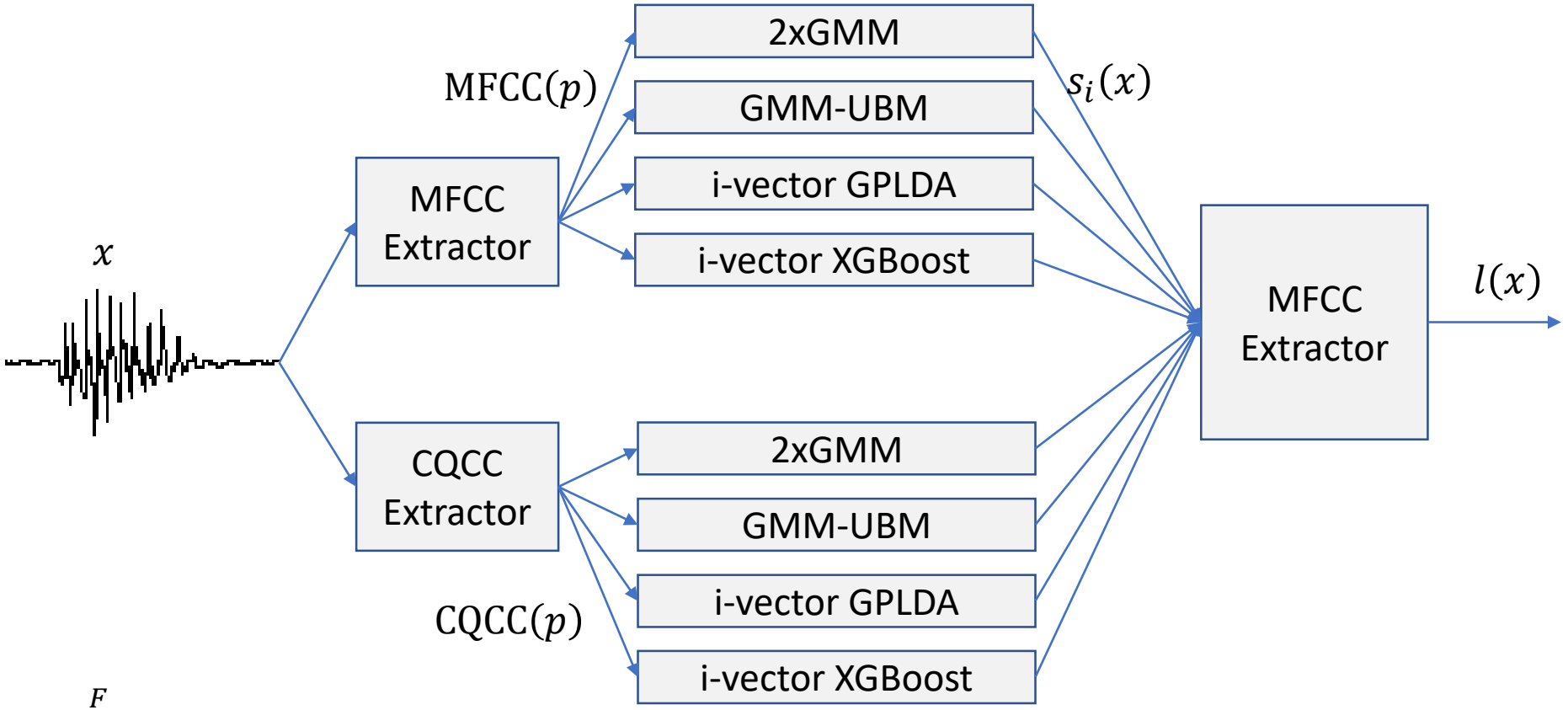
[T. Chen. 2014]

$score(x) = \sum_{i=1}^{K} f_k(x)$, where x is the i-vector for voice sample

ASV Spoof 2017 Dataset

Type of sample	Subset		
	Train	Development	Evaluation
“genuine” speech	1508	760	1298
“replayed”speech	1508	950	12000

Fusion of Classifiers



$$l(x) = a + \sum_{i=1}^{F} b_i s_i(x) + q'Wr$$

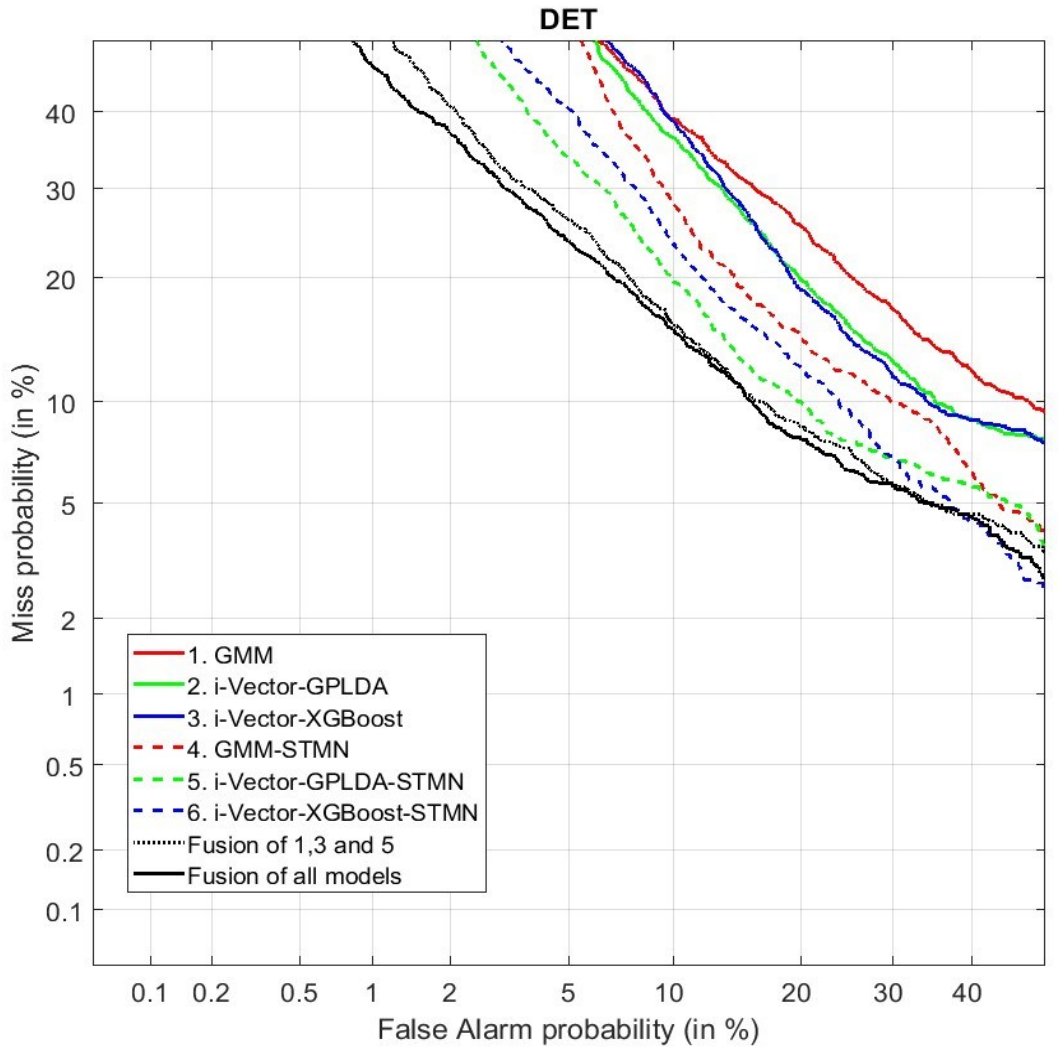
Performance Evaluation

$$P_{fa}(\theta) = \frac{\#\{replay\ trials\ with\ score > \theta\}}{\#\{total\ replay\ trials\}}$$

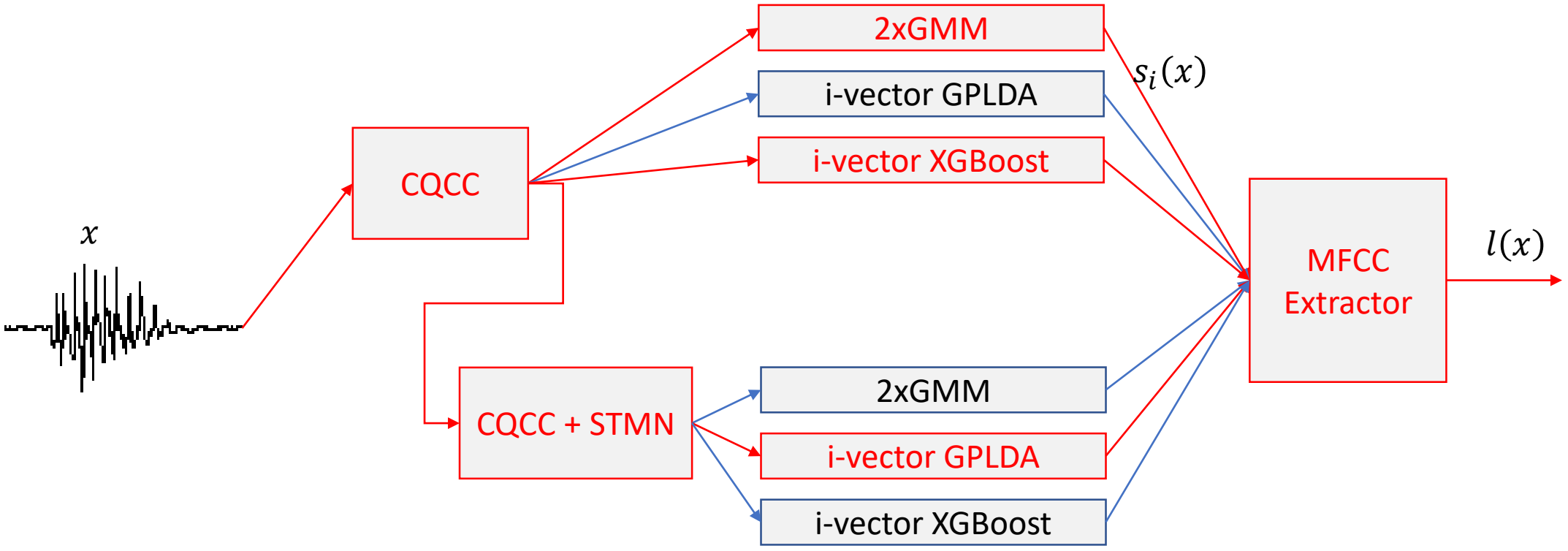
$$P_{miss}(\theta) = \frac{\#\{genuine\ trials\ with\ score < \theta\}}{\#\{total\ genuine\ trials\}}$$

$$P_{fa}(\theta_{EER}) = P_{miss}(\theta_{EER}) = \mathrm{EER}$$

Classifier	EER, %
GMM (Baseline)	22.47
i-vector-GPLDA	20.00
i-vector-XGBoost	19.42
GMM-STMN	16.70
i-vector-GPLDA-STMN	13.81
i-vector-XGBoost-STMN	15.56
Fusion of GMM + i-vector-XGBoost + i-vector-GPLDA-STMN	12.69
Fusion of all models	12.44
1st place in ASV Spoof Challenge 2017, acc. [1]	6.73
2nd place in ASV Spoof Challenge 2017, acc. [1]	12.34
3rd place in ASV Spoof Challenge 2017, acc. [1]	14.03



Fusion of Classifiers



$$l(x) = a + \sum_{i=1}^{F} b_i s_i(x) + q'Wr$$

Conclusion

- Proposed fusion system can provide substantially better performance than the GMM baseline for detection the audio replay attacks
- The normalization of cepstral features is crucial for better performance of replay attack detecting algorithms
- High evaluation performance could be obtained using only few algorithms in a set. The achieved value of EER=12.44% for our fusion classifier is competitive with the best results obtained during ASV Spoof Challenge 2017

This work is supported by the RFBR grant 17-47-220739r_a

ASV Spoof 2017 1st place

Table 1: *CNN architecture*

Type	Filter / Stride	Output	#Params
Conv1	5 × 5 / 1 × 1	864 × 400 × 32	832
MFM1	–	864 × 400 × 16	–
MaxPool1	2 × 2 / 2 × 2	432 × 200 × 16	–
Conv2a	1 × 1 / 1 × 1	432 × 200 × 32	544
MFM2a	–	432 × 200 × 16	–
Conv2b	3 × 3 / 1 × 1	432 × 200 × 48	7.0K
MFM2b	–	432 × 200 × 24	–
MaxPool2	2 × 2 / 2 × 2	216 × 100 × 24	–
Conv3a	1 × 1 / 1 × 1	216 × 100 × 48	1.2K
MFM3a	–	216 × 100 × 32	–
Conv3b	3 × 3 / 1 × 1	216 × 100 × 64	13.9K
MFM3b	–	216 × 100 × 32	–
MaxPool3	2 × 2 / 2 × 2	108 × 50 × 32	–
Conv4a	1 × 1 / 1 × 1	108 × 50 × 64	2.1K
MFM4a	–	108 × 50 × 32	–
Conv4b	3 × 3 / 1 × 1	108 × 50 × 32	9.2K
MFM4b	–	108 × 50 × 16	–
MaxPool4	2 × 2 / 2 × 2	54 × 25 × 16	–
Conv5a	1 × 1 / 1 × 1	54 × 25 × 32	544
MFM5a	–	54 × 25 × 16	–
Conv5b	3 × 3 / 1 × 1	54 × 25 × 32	4.6K
MFM5b	–	54 × 25 × 16	–
MaxPool5	2 × 2 / 2 × 2	27 × 12 × 16	–
FC6	–	32 × 2	332K
MFM6	–	32	–
FC7	–	2	64
Total	–	–	371K

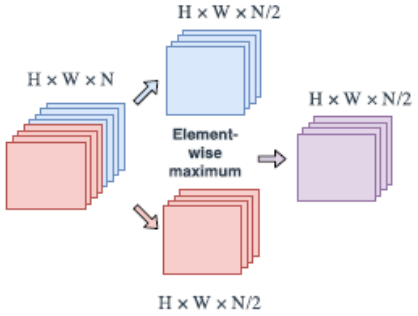


Figure 3: *MFM for convolutional layer*

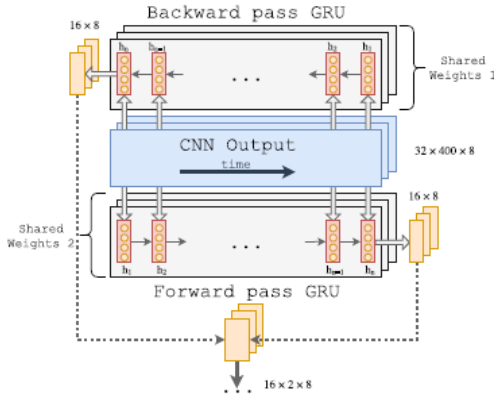


Figure 4: *Bidirectional GRU*

Table 3: *Results on the ASVspoof database*

Individual system	EER (%) Dev dataset	EER (%) Eval dataset
$Baseline$	10.35	30.60
$Baseline_{MVN}$	9.85	17.31
SVM_{i-vect}	9.80	12.54
$LCNN_{FFT}$	**4.53**	**7.37**
$LCNN^{SW}_{FFT}$	5.25	11.81
$LCNN_{CQT}$	4.80	16.54
$CNN_{FFT} + RNN$	7.51	10.69
Fusion system		
$LCNN_{FFT}, SVM_{i-vect}, CNN_{FFT} + RNN$	**3.95**	**6.73**

[G. Lavrentyeva, S. Novoselov et al. 2017]