



# Database Specifications

Machine Vision Group



[www.ee.oulu.fi/mvg](http://www.ee.oulu.fi/mvg)

Avignon, 14/03/2007














# Specifications...?



- Who will record the database?
- When and where to record the database?
- Recording formats?
- Use of different mobile phones?
- Use of different cameras?
- Use of different microphone types?
- Recording inside and outside?
- Considering noisy and clean environment?
- User cooperation? frontal views? different views?
- Facial expressions? Lighting conditions?
- Text dependent or text independent?
- What Speech?
- Which Phrases? Which Digits?
- Which Languages?
- Number of persons in Database? Number of sessions?
- Ethnicity? Ages ? Gender?
- Time intervals between sessions?
- Imposters attacks?
- Database partition? Development, Tuning and Evaluation Sets ?
- Privacy, legal aspects and distribution of the database?
- Other issues?



# Existing Databases

- XM2VTS (Audio-Video)  S. Pigeon et al. The M2VTS multimodal face database (release 1.00), in Proc. 1st Int. Conf. Audio- and Video-Based Biometric Person Authentication, Crans-Montana, Switzerland, 1997, pp. 403–409.
- XM2VTSDB (Audio-Video)  K. Messer et al. XM2VTSDB: The extended M2VTS database, in Proc. 2nd Int. Conf. Audio- and Video-Based Biometric Person Authentication, Washington, DC, 1999, pp. 72–77.
- BANCA (Audio-Video)  E. Bailly-Bailliere et al. The BANCA database and evaluation protocol, in Proc. Audio- and Video-Based Biometric Person Authentication, Guilford, 2003, pp. 625–638.
- VidTIMIT (Audio-Video)  Sanderson et al. Noise compensation in a person verification system using face and multiple speech features, Pattern Recognition, vol. 36, no. 2, pp. 293–302, Feb. 2003.
- VALID (Audio-Video)  Fox et al. The realistic multi-modal VALID database and visual speaker identification comparison experiments, in Lecture Notes in Computer Science, T. Kanade, A. K. Jain, and N. K. Ratha, Eds. New York: Springer-Verlag, 2005, vol. 3546, p. 777.
- BT-DAVID (Audio-Video)  Chibelushi et al. BT DAVID Database Internal Rep., Speech and Image Processing Research Group, Dept. of Electrical and Electronic Engineering, Univ. of Swansea, 1996.
- AVICAR (Audio-Video)  Lee et al. AVICAR: Audio-visual speech corpus in a car environment, in Proc. Conf. Spoken Language, Jeju, Korea, 2004.
- CRIM (Audio-Video)  Audio-Visual French Canadian speech database - Database:  
<http://www.crim.ca/fr/Services/R-D/Vision/dataAccess/dataAccess2/>
- BIOMET (Audio, Video, Signature, Hand, Fingerprint)  Carcia-Salicetti et al. BIOMET: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. AVBPA'03, pp. 845-853.
- BIOSECURE (Audio, Video, Signature, Hand, Fingerprint, Iris)  BIOSECURE database  
<http://www.biosecure.info/>
- SECUREPHONE PDA DATABASE (Audio, Video, Signature)  Koreman et al. multi-modal biometric authentication on the SecurePhone PDA, Proc. Of Second Workshop on Multimodal User Authentication MMUA, 2006.
- ...etc.



# Some Audio-Visual systems



From:  
Aleksic, P.S. & Katsaggelos, A.K.  
**Audio-Visual Biometrics**  
Proceedings of the IEEE (2006)  
Volume: 94, Issue: 11 2025-2044



System	Features		Database	Non-ideal Conditions	Expert	AV Fusion Method	Recognition Mode*
	Acoustic	Visual					
Luetin <i>et al.</i> [135]	none	shape- and appearance-based, and joint (concatenation)	Tulips1	none	HMMs GMMs	none	TD+TI/ID
Chibelushi <i>et al.</i> [5]	MFCCs	shape-based (PCA,LDA, concatenation)	10 speakers [5]	white noise at different SNRs	ANNs	opinion fusion (weighted summation)	TD/ID
Brunelli and Falavigna [6]	MFCCs+ $\Delta^{**}$ + $\Delta\Delta$	appearance-based	89 speakers 3 sessions	none	VQ	opinion fusion (weighted product)	TI/ID
Ben-Yacoub <i>et al.</i> [7]	LPCs	appearance-based	XM2VTS	none	HMMs, sphericity measure [7]	post classifier using binary classifiers (SVM, Bayesian classifier, FLD, decision tree and MLP)	TD+TI/VER
Sanderson and Paliwal [8]	MFCCs+ $\Delta$	appearance-based (PCA)	VidTIMIT	white and operations-room noise at different SNRs	GMMs	weighted summation, concatenation, adaptive weighted summation, SVM, Bayesian classifier	TI/VER
Hazen <i>et al.</i> [9]	MFCCs	appearance-based	35 speakers [9]	data recorded on a handheld device	SVMs	opinion fusion (weighted summation)	TD/ID
Jourlin <i>et al.</i> [10]	LPCs+ $\Delta$ + $\Delta\Delta$	appearance- and shape-based features	M2VTS	none	HMMs	opinion fusion (weighted summation)	TD/VER
Wark <i>et al.</i> [11-13]	MFCCs	shape-based (PCA and LDA)	M2VTS	white noise at different SNRs	GMMs	opinion fusion (weighted summation)	TI/ID+VER
Aleksic and Katsaggelos [14]	MFCCs+ $\Delta$ + $\Delta\Delta$	shape-based (PCA applied on lip-contours)	AMP/CMU	white noise at different SNRs	HMMs	feature-level concatenation	TD/ID+VER
Chaudhari <i>et al.</i> [15]	MFCCs	appearance-based (DCT applied on ROI)	IBM	none	GMMs	feature-level concatenation, opinion fusion	TI/ID+VER
Bengio <i>et al.</i> [32, 33]	MFCCs+ $\Delta$	shape-based and appearance-based	M2VTS	white noise at different SNRs	asynchronous HMMs	midst-mapping fusion	TD /VER
Fox <i>et al.</i> [34,35]	MFCCs+ $\Delta$	appearance-based (DCT)	XM2VTS	white noise at different SNRs	HMMs	feature-level concatenation, opinion fusion (weighted summation)	TD/ID
Nefian <i>et al.</i> [30]	MFCCs+ $\Delta$ + $\Delta\Delta$	appearance-based (PCA+LDA)	XM2VTS	white noise at different SNRs	Coupled HMMs embedded HMMs	midst-mapping fusion, opinion fusion (weighted summation)	TD/ID
Kanak <i>et al.</i> [38]	MFCCs+ $\Delta$ + $\Delta\Delta$	appearance-based (PCA)	38 speakers [38]	white noise at different SNRs	HMMs	concatenation, opinion fusion (Bayesian fusion)	TD/ID

\* TD: text-dependent; TI: text-independent  
VER: verification; ID: identification

\*\*  $\Delta$  – first derivative  
 $\Delta\Delta$  – second derivative

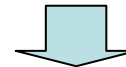


## **XM2VTSDB (Audio-Video)**

## XM2VTSDB (Audio-Video)



- 295 subjects !! ( $360-65=295$ )
- Blue uniform background, controlled illuminations, fixed camera settings.
- Audio recorded at sampling rate of 16bit & frequency of 32 khz
- 4 Sessions, 1 month interval between two sessions!!
- In each session, for each subject  $\rightarrow 2 \times (3 \text{ speech shots} + 1 \text{ head rotation shot})$

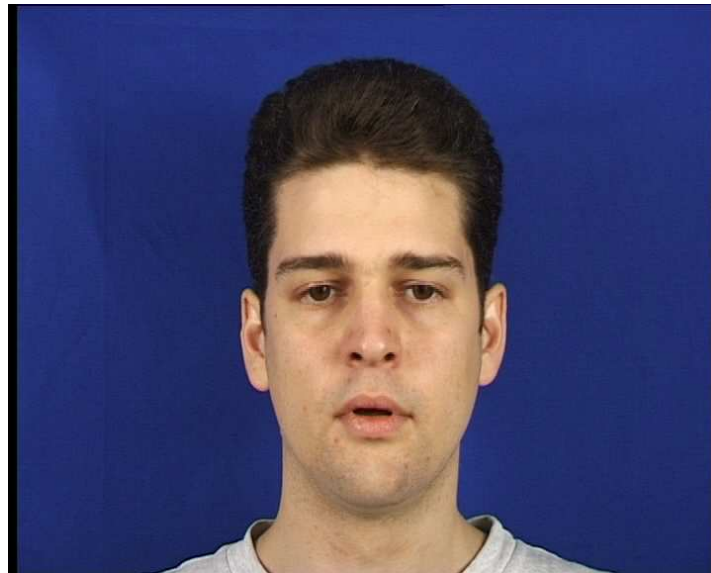


Utterances of 3 fixed phrases:

{ zero one two three four five six seven eight nine "  
 { five zero six nine two eight one three seven four "  
 { Joe took fathers green shoe bench out "







200

+

25

+

70

Session	Shot	Clients	Impostors	
1	1	Training	Evaluation	
	2	Evaluation		
2	1	Training		
	2	Evaluation		
3	1	Training		
	2	Evaluation		
4	1	Test		Test
	2			

(a) Configuration I

200

+

25

+

70

Session	Shot	Clients	Impostors	
1	1	Training	Evaluation	Test
	2			
2	1			
	2			
3	1	Evaluation		
	2			
4	1	Test		
	2			

(b) Configuration II

Figure 1: The partitioning of the extended M2VTS database according to configuration (a) I and (b) II of the protocol.



## **BANCA (Audio-Video)**

































## BANCA (Audio-Video)



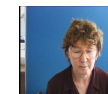
- 208 subjects ( $\frac{1}{2}$  men,  $\frac{1}{2}$  women)
- 12 sessions, during 3 months.
- 4 European languages (English<sup>(52)</sup>, French<sup>(52)</sup>, Italian<sup>(52)</sup> & Spanish<sup>(52)</sup>)
- 3 Scenarios: controlled, degraded and adverse.
- Both high and low quality microphones (2 simultaneous) and cameras (2) were used.
- The subjects were asked to say a random 12-digit number, their name, their address and date of birth during each recording.
- Each recording lasts about 20 seconds. Audio recorded at both 16bit & 32bit at 32 khz



## 208 Subjects

English 52				French 52				Italian 52				Spanish 52			
															
26 m 		26 f 		26 m 		26 f 		26 m 		26 f 		26 m 		26 f 	
g1=13 	g2=13 	g1=13 	g2=13 	g1=13 	g2=13 	g1=13 	g2=13 	g1=13 	g2=13 	g1=13 	g2=13 	g1=13 	g2=13 	g1=13 	g2=13 

Controlled	Session 1
	Session 2
	Session 3
	Session 4
Degraded	Session 5
	Session 6
	Session 7
	Session 8
Adverse	Session 9
	Session 10
	Session 11
	Session 12



True Client	Impostor Attack
0 3 8 9 2 1 6 7 4 5 0 1	8 5 7 9 0 1 3 2 4 6 0 2
Annie Other	Gertrude Smith
9 St Peters Street	12 Church Road
Guildford	Portsmouth
Surrey	Hampshire
GU2 4TH	PO1 3EF
20.02.1971	12.02.1976

**Table 1:** Example of the speech uttered by a subject at one of the twelve Banca sessions.



**BIOSECURE Database**  
(Audio, Video, Signature, Hand, Fingerprint, IRIS)

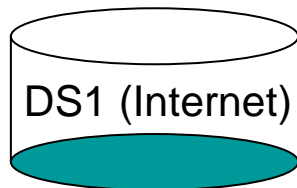


FP6 Project (2004-2007)

# BIOSECURE (Audio, Video, Signature, Hand, Fingerprint, IRIS) Biometrics for Secure Authentication

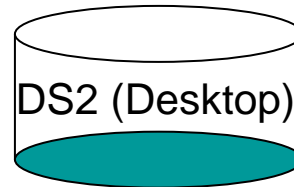


Samsung Q1 - Celeron M 900 MHz  
ULV - UMPC - RAM 512 Mo - HDD 40  
Go - GMA 900 - LAN sans fil :  
802.11b/g, Bluetooth 2.0 EDR - Win  
XP Tablet PC - 7" écran large TFT 800  
x 480 ( WVGA )



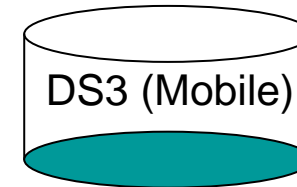
Voice, face

PC-based  
On-line  
Unsupervised  
(Internet)  
~1000 Subjects ?



Voice, face, signature, fingerprint,  
hand, iris

PC-based,  
Offline, Supervised  
~700 Subjects ?  
2 Sessions



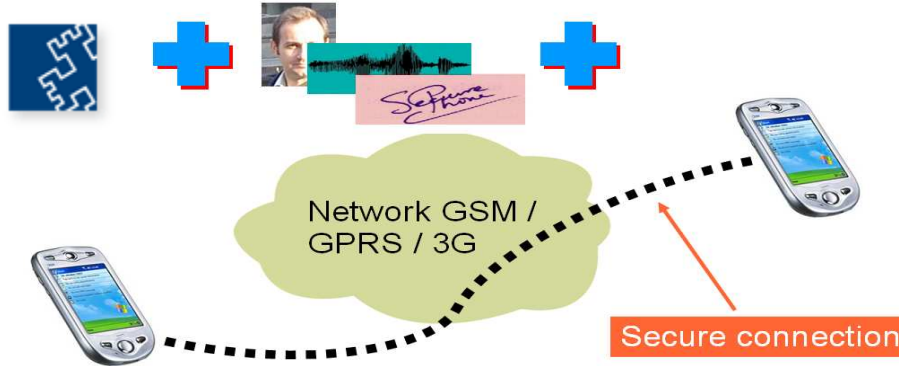
Voice, face, signature,  
fingerprint

PDA-based,  
indoor/outdoor  
~700 Subjects?  
2 Sessions



## SecurePhone PDA Database

<http://www.secure-phone.info/>



## SecurePhone consortium

securephone

- » SecurePhone is a research project funded partly by the European Commission's Framework Program 6
- » Project developed by a consortium of 7 partners:



Atos Origin: Project Coordinator



The University of Buckingham



Informa



Nergal



Telefónica Móviles  
España



UNIVERSITÄT  
DES  
SAARLANDES

Universität des Saarlandes





## SecurePhone PDA Database

<http://www.secure-phone.info/>



- 3 Modalities (Audio, Video, Signature)
- 60 subjects (30 females, 30 males)
- Data recorded on PDA (QTEK2020 / XDA II)
- 5 digits, 10 digits + Short phrases
- Recorded in quiet and noisy environments both inside (office) and outside (street)
- 3 sessions, separated at least by 1 week interval
- Each session: 2 indoor (light-clean & dark-noisy) and 2 outdoor (light-noisy & dark-noisy) recording
- Voice data at 22 kHz, Video Frame at 19.6 fps
- Variable background and lighting conditions
- It is assumed cooperative subjects → face size/pose limited → fit inside a box area on the PDA display.

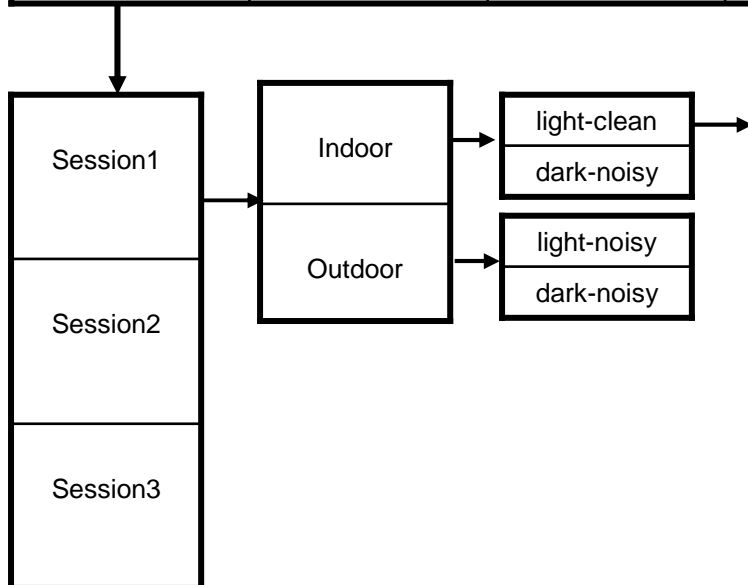
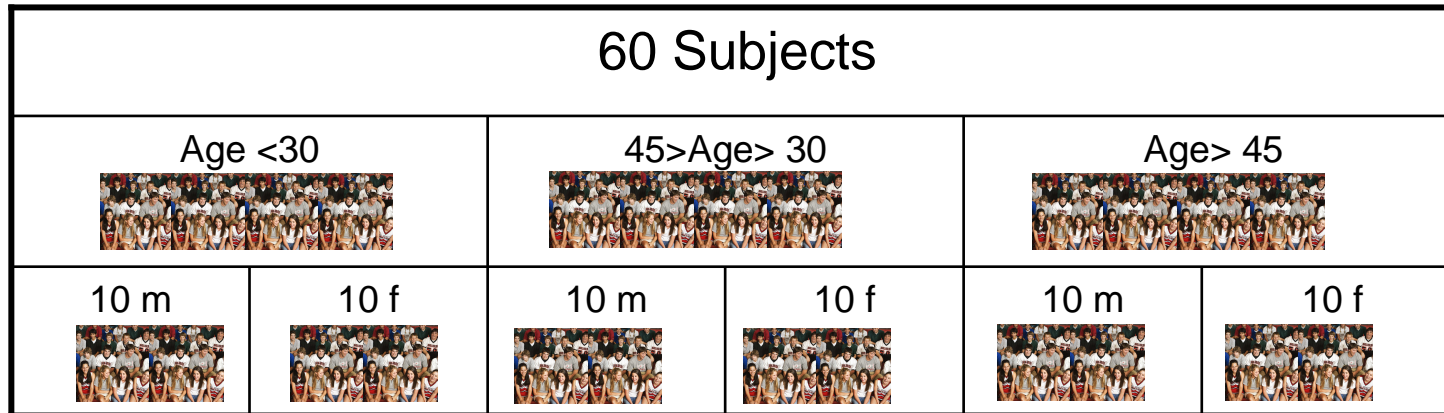
**Performance:** Intel XScale PXA 263 proprocessor running at 400 MHz. 128 megs of RAM and 14 megs of flash file storage. 64 megs ROM. Windows Mobile 2003 Phone Edition operating system.





## SecurePhone PDA Database

<http://www.secure-phone.info/>



Index	5-digit strings
01	5 3 8 2 4
02	6 2 1 9 7
03	4 2 7 1 3
04	2 8 3 7 6
05	1 9 8 5 4
06	4 5 2 3 9
Index	10-digit strings
07	4 3 1 3 8 7 4 6 1 5
08	2 9 2 8 7 3 7 9 3 8
09	5 7 9 2 4 7 9 1 2 6
10	3 9 6 4 6 3 7 6 3 1
11	6 4 2 1 4 7 1 5 3 4
12	1 2 6 1 6 9 2 9 8 1
Index	phrases
01	Stop each car if its little
02	Play in the street up ahead
03	A fifth wheel caught speeding
04	Charlie, did you think to measure the tree?
05	Tina got cued to make a quicker escape
06	Here I was in Miami and Illinois

60 Subjects x 3 Sessions x 2 locations (indoor, outdoor) x 2 recording conditions x 3 prompts types x 6 examples of prompt types = **12960 recording**.

- » **Corporate:**
  - » Secure transactions between sales force and headquarters
  - » Secure transactions with suppliers
- » **Commercial:**
  - » Mobile electronic banking
  - » Mobile private banking
  - » Mobile payments, on-line reservations
- » **Governments and Public Administration:**
  - » Administrative transactions through mobile phone
  - » Police force communications
  - » Mobile medical services





# More Databases ..

	<b>VidTIMIT</b> ↓	<b>VALID</b> ↓	<b>DAVID</b> ↓	<b>CRIM</b> ↓	<b>AVICAR</b> ↓	<b>BIOMET</b> ↓
Modalities	Audio-Video	Audio-Video	Audio-Video	Audio-Video	Audio-Video	Audio, Video, Signature, Hand, Fingerprint
Number of Subjects	<b>43</b> (24 M + 19 F)	<b>106</b> (77 M + 29 F)	<b>&gt; 100</b>	<b>20</b>	<b>100</b> (50 M + 50 F)	<b>131</b> (106) ( 92 )
Number of Sessions	<b>3</b>	<b>5</b>	<b>5</b> (30 subjects)	-	-	<b>3</b>
Interval	1 week	In 1 month	Months	-	-	3 & 5 months
Utterances	10 phonetically balanced sentences	Same as in XM2VTS	Digits, alphabet phrases	Reading broadcast news (text independent)	Isolated digits, letters, phone numbers, sentences	PIN code 0→9, 9→0 "Oui", "Non" 12 balanced sentences
Environment , Comments	Head to left, right, up and down	Acoustic noise, illumination changes, head rotation	Frontal and profile views	591 sequences 5 hours Between 23-47 videos per person	Inside a car Five car noise conditions Different car speeds	Frontal, ±15° and ±45° 1 sequence~90s



# Some Suggestions!!



- Who will record the database?
- When and where to record the database?
- Recording formats?
- Use of different mobile phones?
- Use of different cameras?
- Use of different microphone types?
- Recording inside and outside?
- Considering noisy and clean environment?
- User cooperation? frontal views? different views?
- Facial expressions? Lighting conditions?
- Text dependent or text independent?
- What Speech?
- Which Phrases? Which Digits?
- Which Languages?
- Number of persons in Database? Number of sessions?
- Ethnicity? Ages ? Gender?
- Time intervals between sessions?
- Imposters attacks?
- Database partition? Development, Tuning and Evaluation Sets ?
- Privacy, legal aspects and distribution of the database?
- Other issues?



# Some Suggestions!!



- Who will record the database? B. Crettol (IDIAP) ☺ with local help
- When and where to record the database?  
at all sites, Sept 08 → March 08 , Each site should find “many” volunteers and ensure that probably they will participate in different sessions ☺  
IDIAP (30), UNIS (30), UMAN (30), OULU (20), LIA(20 ), BUT (30) → 160 persons  
(30 % should be females)
- Recording formats? Giorgio ☺
- Use of different mobile phones? “mainly” record using 1 type of mobile phones ☹
- Use of different cameras?
- Use of different microphone types?
- Recording inside and outside? Possibly both (according to the chosen scenarios) !!
- Considering noisy and clean environment? Possibly both (with use of headsets) !!
- User cooperation? frontal views? different views? Only near frontal views!!
- Facial expressions? Lighting conditions?  
facial expressions not addressed, uncontrolled illumination.
- Text dependent or text independent? Dialogue oriented.





# Some Suggestions!!



- What Speech? Dialogue to be defined after scenarios
- Which Languages? English
- Number of persons in Database? 160
- Number of sessions? As many sessions as possible, one Session=one day, 1 shot =1 repetition.
- Ethnics? Ages ? Gender?  
Cross European diversity (faces etc.). Different categories of ages (18-65). Gender Balanced ( $\pm 30\%$ )!! Statistics and info about the subjects (gender, mother tongue etc.) should also be collected for internal use.
- Imposters attacks? e.g.:
  - Stolen code
  - Picture/Video
  - Recorded Speech
  - Picture (Video) + Recorded Speech(..this issue will be considered during the second recording phase)



# Some Suggestions!!



- Database partition? Development, Tuning and Evaluation Sets ? e.g. Like in XM2VTS, BANCA OR DEFINE OWN PROTOCOL ..  
(this issue will be developed later ..)
- Privacy, legal aspects and distribution of the database? An agreement to be signed by the participants about the public use of the data, distribution, inclusion in publications, privacy etc. IDIAP will prepare the agreement proposal. We only consider subjects who accept that data will be distributed.

- Additional datasets?

Optionally, additional small datasets could also be collected:

1) Data recording using different mobile phones.

2) Data under very realistic conditions: The idea is to give (borrow ☺) a number of mobile phones e.g. for a period of two months to certain persons and ask them to regularly record data at home, in the streets or wherever they want → Very natural conditions. Then the data will be collected and the mobile phones given to other persons and so on.





# Some Suggestions!!

## MoBio Database



IDIAP 30	UMAN. 30	UNIS 30	LIA 20	BUT 30	Oulu 20	EPM 0	IDEA 0
-------------	-------------	------------	-----------	-----------	------------	----------	-----------

