Improving Robustness of Speaker Verification by Fusion of Prompted Text-Dependent and Text-Independent Operation Modalities

Iosif Mporas, Saeid Safavi, and Reza Sotudeh

Information Engineering and Processing Architectures Group
Electronics Communications and Electric Division
School of Engineering and Technology
University of Hertfordshire
College Lane Campus, Hatfield AL10 9AB, UK

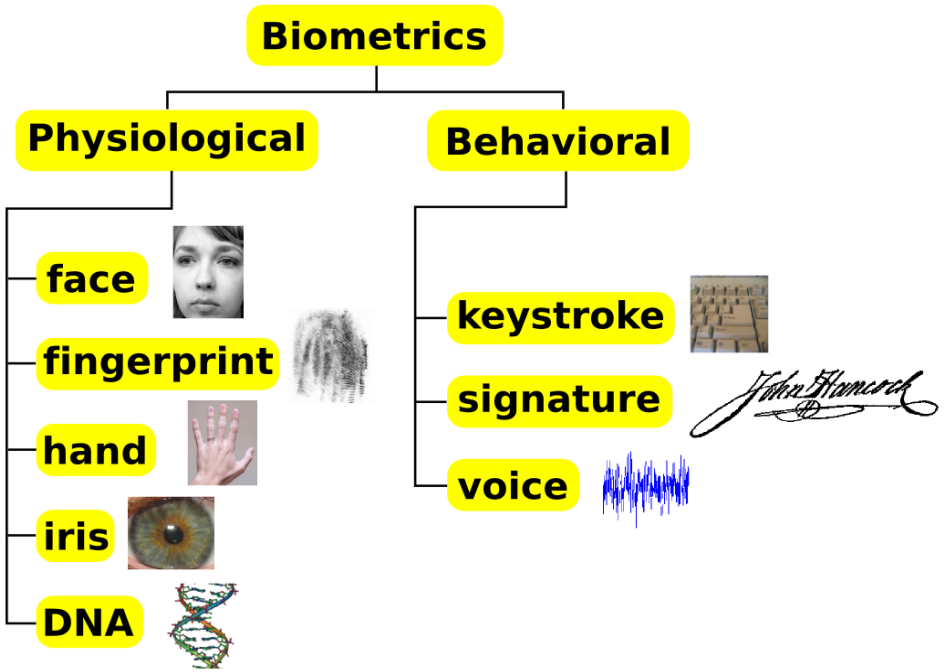
OCTAVE
Objective Control of Talker Verification

Outline

- Introduction
- Speaker Verification
- SV Methodology
- Experimental Setup
- Experimental Results
- Conclusion

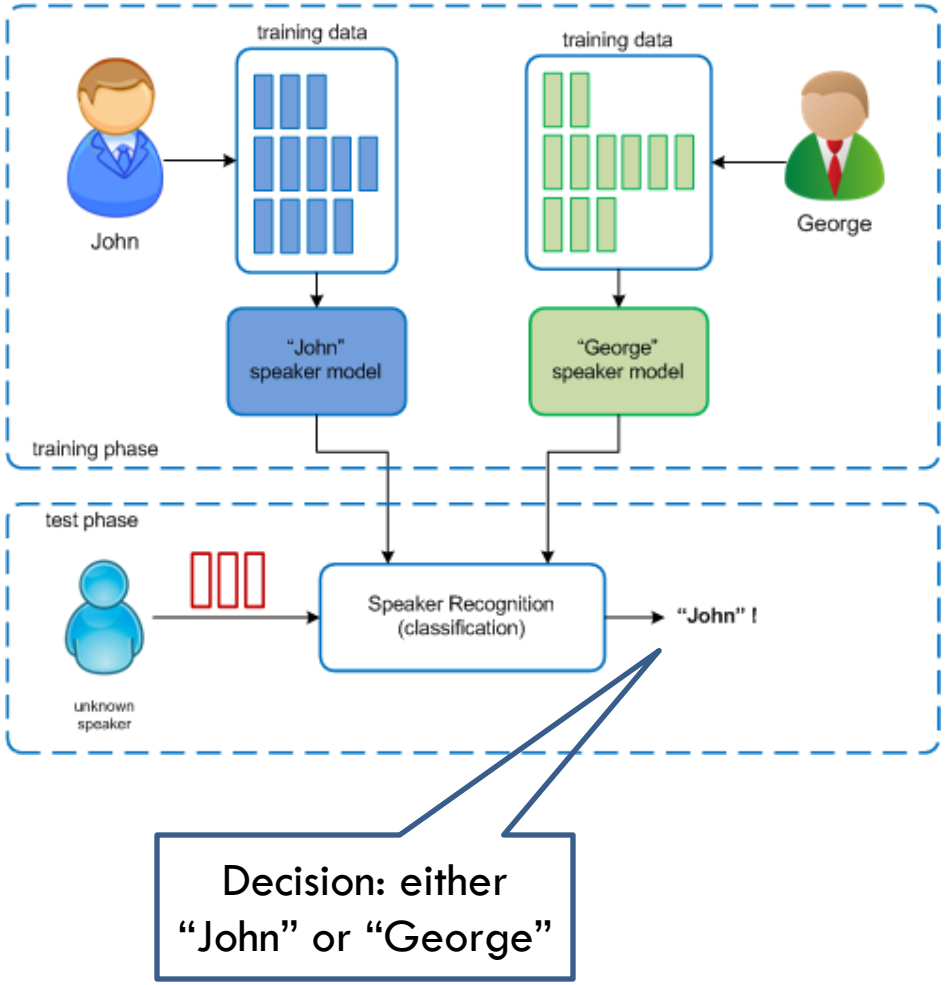
Introduction

- Biometrics and Voice
  - security access control to physical places
  - secure login to computer systems and mobile devices
  - online banking
  - personalized human-machine interfaces
- Voice biometrics
  - offer convenience to the users
  - rely on microphones not on special sensors

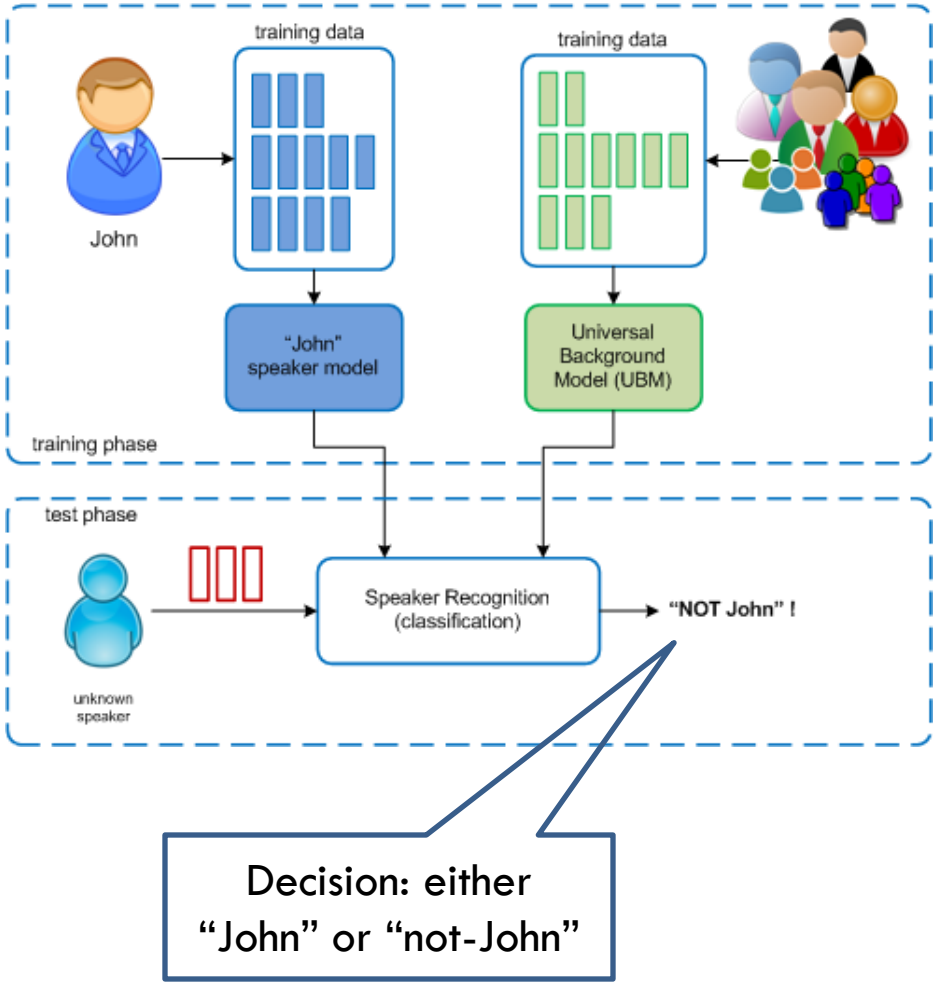


Speaker Recognition

Speaker Identification



Speaker Verification



Speaker Verification

- text-dependent speaker verification
  - users pronounce a pre-determined pass-phrase
  - The pass-phrases are either unique or prompted by the system, e.g. in a screen.
  - High SV performance
  - Vulnerable to spoofing attacks
- text-independent speaker verification
  - the speech content is not apriori known
  - Lower SV performance
  - Robust to spoofing attacks

Mode of operation		*Recognition accuracy*	*Convenience for clients*	*Spoofing*
Passphrase based voice biometrics (TD)		High	High	Easy
Text prompted voice biometrics	Text-dependent (TD)	High	Medium	Medium
	Text-independent (TI)	Medium	Low	Hard
Intrinsic TI voice biometrics, i.e. verify the client with speech uttered for interaction in automated speech driven systems		Medium	Training: Low	Medium
			Testing: High	

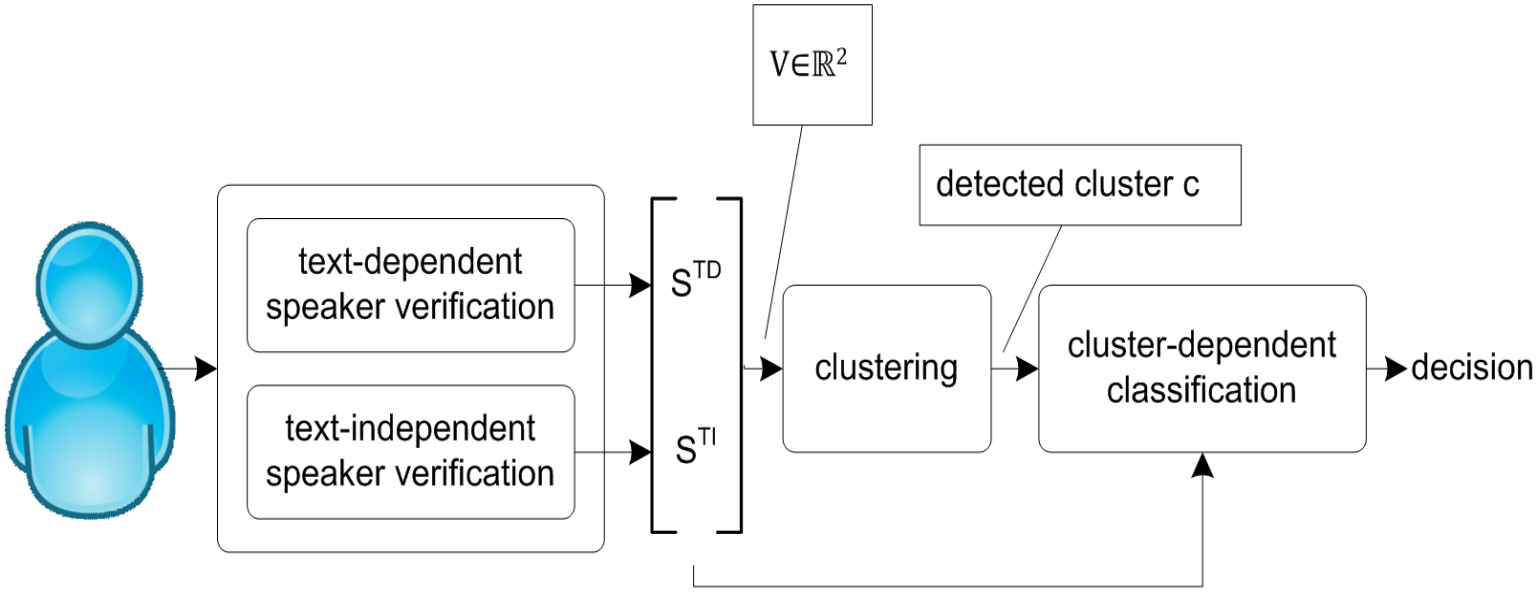
Speaker Verification technology

- short-time analysis of the voice signal and post-processing by a pattern recognition algorithm
- Speech descriptors (features)
  - Mel frequency cepstral coefficients (MFCCs)
- Modeling of speakers
  - Gaussian Mixture Models (GMMs)
    - GMM-UBM
    - maximum a-posteriori (MAP) adaptation or means-only adaptation of the UBM to speaker specific data
- support vector machines (SVMs)
  - in combination with GMMs by concatenating the means of the Gaussian components of the GMMs to super-vectors and afterwards apply discriminative classification on them
- Factor analysis
  - i-vectors

- In specific experimental setups, i-vector method has outperformed the classic GMM-UBM approach
- However, GMM-UBM based modeling offers more stable results, with respect to the availability of significantly large amount of training data or not

Fusion of Speaker Verification Operation Modes

- Fusion of the prompted text-dependent (TD) and text-independent (TI) modes of operation is performed on score level
- Since the score values typically present some variations, in order to support the classification stage, we apply in advance clustering to separate the 2-dimentional score data to areas with less variation.
- After clustering the data we apply a cluster-specific classification model and get the verification decision.



Experimental Setup

University of Hertfordshire

- Speech Data
  - RSR2015 speech corpus
    - recordings from 300 speakers (157 males, 143 females)
    - for each speaker, there are 3 enrolment sessions of 73 utterances each and 6 test sessions of 73 utterances each
    - in total there are 657 utterances distributed in 9 sessions per each speaker
    - 16 kHz, 16 bits
  - TIMIT
    - for training a universal background model
    - recordings of 630 speakers
    - 16 kHz, 16 bits

Experimental Setup

- pre-processing and parameterization
  - energy-based speech activity detector
  - framing: time shifting Hamming window (20 msec) time shift between successive frames (10 msec)
  - first 19 MFCCs + Δ + ΔΔ (d = 57)
  - In order to reduce the effect of handset mismatch and make the features more robust, RASTA and CMVN were applied to the MFCC features.
- Speaker verification models
  - GMM-UBM approach
  - UBM was built by a mixture of 128 Gaussian distributions and was trained using all utterances from 630 speakers from TIMIT.
  - For each of the speakers of the RSR2015 database we applied means only adaptation on the UBM model, using the speaker-specific enrollment data.

Experimental Setup

- Evaluation protocol
  - Text-dependent operational scenario
  - Text-independent operational scenario

- Classification algorithms (fusion)
  - multilayer perceptron neural networks (MLP)
  - C4.5 decision trees (C4.5)
  - support vector machines (SVM)
  - Bayesian networks (BN)
  - classification and regression trees (CART)
  - reduced error pruning tree (REP)
- For the implementation of these machine learning algorithms for classification we used the WEKA toolkit

Experimental Results (1/2)

- 10-fold cross validation protocol

- Speaker verification, in terms of percentages of sensitivity and specificity, for different operation mode fusion methods.

Method	sensitivity	specificity
TD (single mode)	84.65	97.46
TI (single mode)	84.70	91.83
MLP	71.14	99.82
C4.5	72.89	99.79
SVM	71.12	99.82
Bayesian Network	76.33	99.66
CART	72.13	99.81
REP	73.10	99.78

Experimental Results (2/2)

- Speaker verification, in terms of percentages of sensitivity (sens) and specificity (spec), for different operation mode fusion methods and different number of clusters.

	c=1		c=5		c=10		c=20	
Method	sens	spec	sens	spec	sens	spec	sens	spec
TD (single mode)	84.65	97.46	-	-	-	-	-	-
TI (single mode)	84.70	91.83	-	-	-	-	-	-
MLP	71.14	99.82	73.30	99.79	73.30	99.78	72.11	99.81
C4.5	72.89	99.79	72.97	99.79	71.92	99.81	71.53	99.82
SVM	71.12	99.82	76.07	99.70	58.89	99.97	68.43	99.88
Bayesian Network	76.33	99.66	78.70	99.57	86.55	98.88	86.47	98.92
CART	72.13	99.81	73.18	99.79	72.37	99.80	72.47	99.81
REP	73.10	99.78	73.33	99.77	72.52	99.78	72.34	99.79

Conclusion

- We presented a fusion methodology for combining prompted text-dependent and text-independent speaker verification operation modalities
- In order to improve the performance we applied clustering of the score-based data before the classification stage
- The experimental evaluation using clustering of the single mode score data followed by application of classification for fusion showed an absolute improvement of more approximately 2% in terms of sensitivity and an absolute improvement of 1.5% in terms of specificity.
- The best performing algorithm for fusing the two modes of speaker verification operation was found to be the Bayesian network classifier.
- The improvement is owed to the exploitation of the underlying and complementary information between the distributions of the scores of the two modes of operation.
- We deem the fuse of the two modalities can lead to real-world voice biometrics based applications which will be more accurate and thus more robust to spoofing attacks.

Thank you for your attention!

The research reported in the present paper was partially supported by the H2020 OCTAVE Project ("Objective Control for TAlker VErification" - Grand Agreement number 647850) funded by the European Commission under the Horizon 2020 Programme. Project web-site: https://www.octave-project.eu//.