The CRSS Systems for NIST SRE 2010

Yun Lei, Taufiq Hasan, Jun-Won Suh, Abhijeet Sangwan, Hynek Boril, Gang Liu, Keith Godin, Chi Zhang and John H. L. Hansen

email: {taufiq.hasan, john.hansen}@utdallas.edu

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

NIST
Speaker Recognition Evaluation

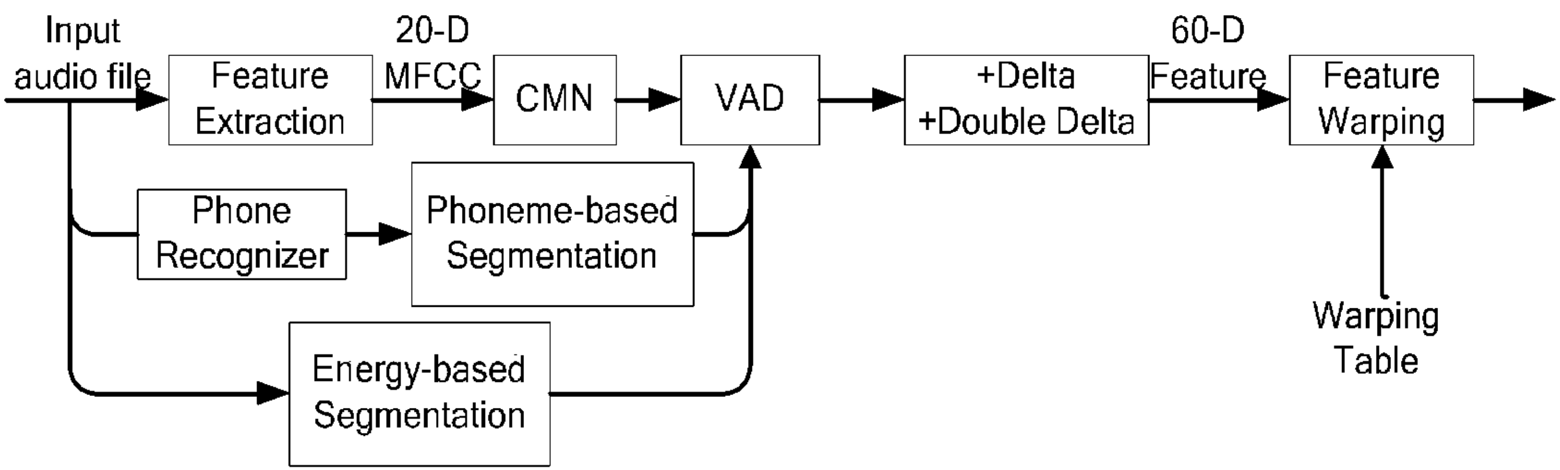
Odyssey 2010
June 24-25, Brno, Czech Republic

Introduction

- **Systems Summary**
  - The CRSS system is a fusion of five SVM based systems [1] and one Joint Factor analysis system [3]
  - The factor analysis based front end [1] is used as features for the SVM based systems
- **Task focus**
  - We mainly focused on the core-core telephone train and test condition
  - We also submitted a system for the 10sec-10sec condition
- **Novel Elements**
  - New background selection strategy was employed
  - Supervised Probabilistic Principal Component Analysis method was introduced

Feature Extraction

- **Algorithm Details**
  - 60-dimension feature (19 MFCC with log energy + Δ+ Δ Δ) using a 25 ms analysis window with 10 ms shift
  - Used feature warping with a 3-s sliding window
  - Used Hungarian phoneme recognizer [6] and simple energy based voice activity detection (VAD)
  - This is the common acoustic front-end for all subsytems



System Components

- **UBM Training**
  - Gender dependent UBMs with 1024 mixtures
  - NIST 2004, 2005, 2006 SRE data used for training
  - 20 iterations per mixture split (HTK toolkit)
- **Factor Analysis (PPCA and SPPCA)**
  - Two different modeling approaches used:
    - Standard Probabilistic principal component analysis (PPCA) [2]
    - New technique: Supervised probabilistic principal component analysis (SPPCA) [4]
  - Data: Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST 2004, 2005, 2006 SRE enrollment data
  - Total 400 factors used

System Components

- **Channel Compensation**
  - Three techniques are used:
    - Linear Discriminant Analysis (LDA)
    - Nuisance Attribute Projection (NAP)
    - Within Class Covariance Normalization (WCCN)
  - Training Data: NIST 2004, 2005, 2006 SRE enrollment data used for training the LDA, NAP and WCCN matrices
- **SVM Training (SVM)**
  - The cosine kernel was used for SVM.
  - Background dataset consists of NIST SRE 2004, 2005, 2006, and the Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, with a total of 12,763 utterances.
  - Used only SRE 04 and 05 as background dataset for final submission

Impostor Selection

- **Proposed Method**
  - The idea is to find the best group of impostor speakers for enrollment speakers [4]
  - Used SVM ranking algorithm to find the closest background set
  - Used SVM-delta for selecting best background set for each enrollment speaker

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x_i} \qquad SVWeight_i = \sum_{k=1}^{n} \alpha_{ik}$$

$$SVMdelta_n = SVWeight_l - SVWeight_m \quad (l < m)$$

NIST

UTD

Score Normalization and Fusion

- **Score Normalization**
  - NIST SRE 2005 data was used for T-norm
  - The T-norm model is trained with a leave-one-out method
  - No Z-norm was used in the SVM systems
- **Score Fusion**
  - Score fusion software based on Brummer et. al.'s FoCal toolkit was implemented [7]
  - Linear logistic regression (LLR) method is used to train the fusion weights
  - The score fusion software is designed to automate the process of choosing a fusion method for the best MinDCF value

The Subsystems

- **SVM Based Subsystems:**
  - SVM-SPPCA-LDA
  - SVM-PPCA-LDA
  - SVM-SPPCA-NAP
  - SVM-PPCA-NAP
  - SVM-PPCA-LDA-BG
  - GMM-UBM-JFA

- **Commonalities**
  - All SVM systems utilize WCCN after LDA or NAP
  - Only the system SVM-PPCA-LDA-BG uses the new background selection algorithm [5]

Joint Factor Analysis Subsystem

- **Subsystem Details**
  - 300 speaker factors and 100 channel factors was used
  - Training data:
    - Eigenvoice Matrix V: Switchboard II, Phases 2 and 3, Switchboard Cellular, Part 1 and 2; NIST 2005 and 2006 data
    - Eigenchannel Matrix U: NIST 2004, 2005, and 2006 data
    - Diagonal Matrix D: NIST 2004 data
  - No score normalization was used in this case
  - Notated as GMM-UBM-JFA in subsequent slides

Fusion

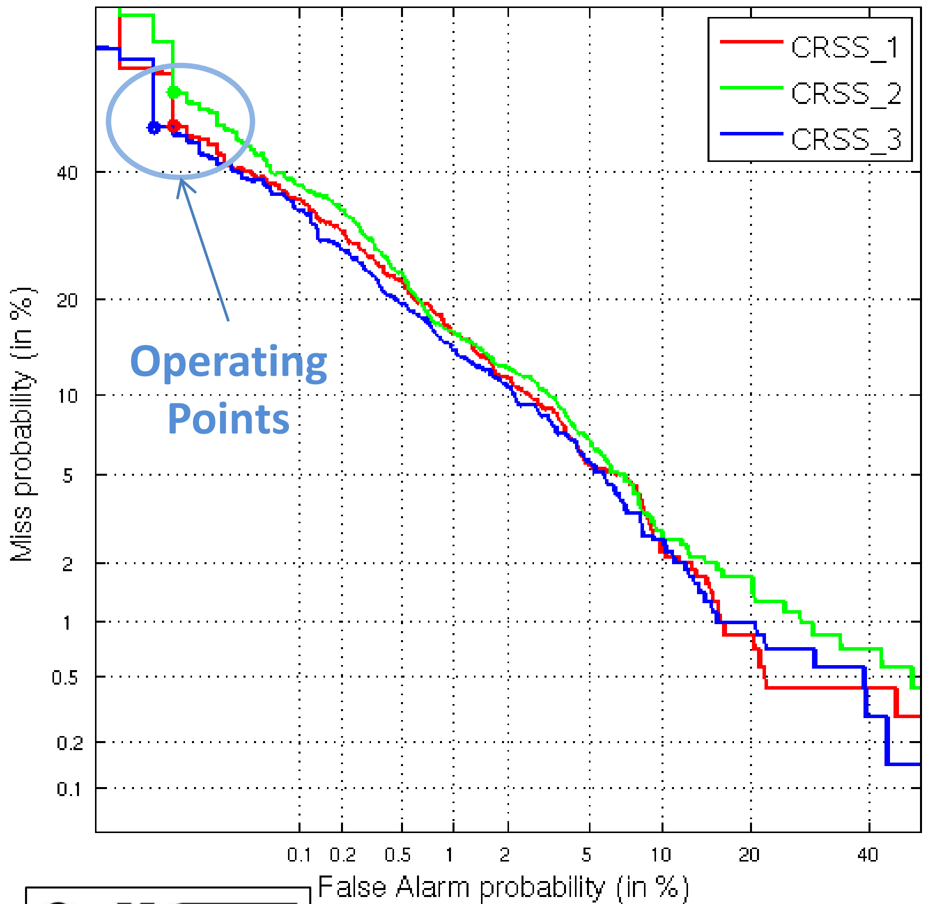
- **Construction of the CRSS Submissions**

Submission	Fused Subsystems
CRSS_1	SVM-PPCA-LDA
	SVM-PPCA-NAP
CRSS_2	SVM-PPCA-LDA
	SVM-PPCA-NAP
	SVM-SPPCA-LDA
	SVM-SPPCA-NAP
	GMM-UBM-JFA
	SVM-PPCA-LDA-BG

Submission	Fused Subsystems
CRSS_3	SVM-PPCA-LDA
	SVM-PPCA-NAP
	SVM-SPPCA-LDA
	SVM-SPPCA-NAP
	SVM-PPCA-LDA-BG

Results

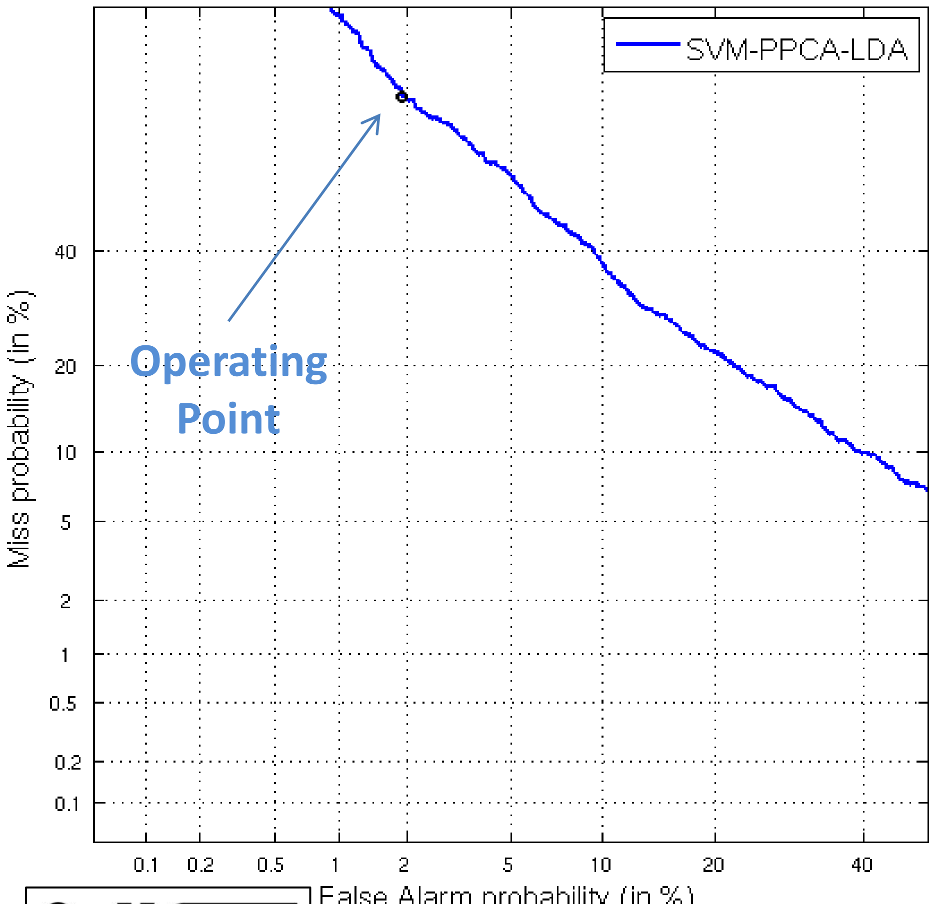
- **Submission Performance** (NIST 2010 SRE, core-core, Cond. 5)



Submission	EER (%)	MinDCF
CRSS_1	**5.225501**	0.585491
CRSS_2	5.791149	0.646226
CRSS_3	5.264267	**0.546166**

Results

- **Submission Performance** (NIST 2010 SRE, 10sec-10sec)

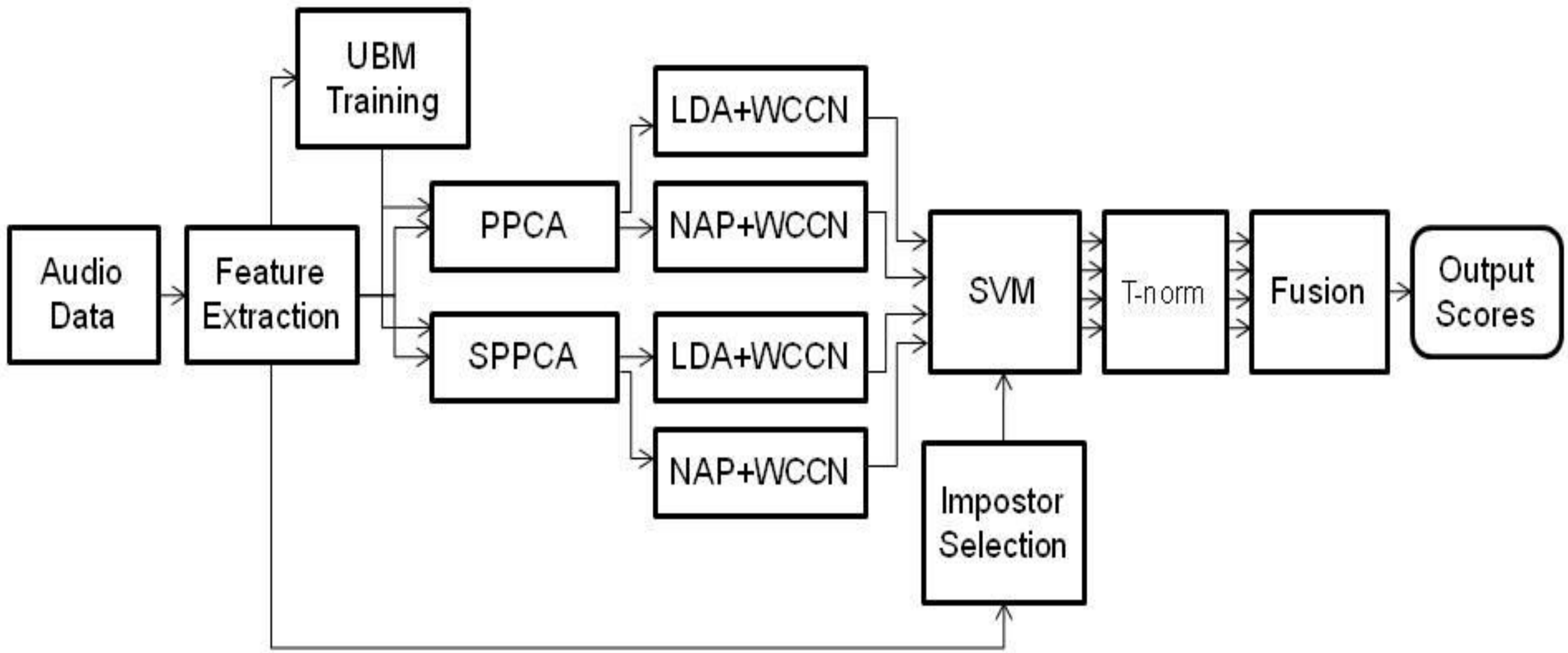


System	EER (%)	MinDCF
SVM-PPCA-LDA	21.119471	0.89685

- We used the SVM-PPCA-LDA system for 10sec case
- Paramater Tuning can further improve the performance

System Block Diagram

- **CRSS SVM Submission Architecture**



Other Developments

- **ASR MLLR System**
  - ASR trained on Switchboard is used to generate MLLR transform matrices for speaker verification tokens
  - The ASR employs PLP front-end and feature warping
  - A global MLLR transform and broad phone-group transforms are estimated
  - PCA is applied to reduce feature dimension and SVM is used as classifier
  - Achieved 21.46% EER for SRE08 core tel-tel for male trials.
- **PMVDR Features Based System**
  - A GMM-UBM-MAP system was evaluated
  - Achieved 13.103% EER for SRE08 core tel-tel for male trials.
  - Requires further investigation

Computational Resources

- **Computational Resources**
  - **System OS:** High performance Rocks computing cluster running the CentOS Linux distribution
  - **CPU:** The cluster comprises 18 HP Intel Quad-Core Xeon 2.33 GHz CPU's. Total 72 CPU cores
  - **RAM:** 126 GB
  - **Disks:** A 4 TB external RAID disk array is used
- **CPU Execution Times**
  - **Training:** Requires 6.2771 mins for a 5 min utterance assuming a single CPU. Real time factor (RTF) = 1.2554
  - **Testing:** Requires 4.6034 mins for a 5 min utterance assuming a single CPU. Real time factor (RTF) = 0.9207

References

[1] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-end Factor Analysis for Speaker Verification," *submitted to IEEE Transaction on Audio, Speech and Language Processing.*

[2] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural computation, vol. 11, no. 2, pp. 443–482, 1999.*

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 4, pp.* 1435–1447, May 2007.

[4] Y. Lei and J. H. L. Hansen, "Speaker recognition using supervised probabilistic principal component analysis," in Proc. Interspeech'10 (Submitted), 2010

[5] J. Suh, Y. Lei, and J. H. L. Hansen, "Best background data selection in SVM speaker recognition for new diverse evaluation data sets," in Proc. Interspeech'10 (Submitted), 2010.

[6] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in Proc. IEEE ICASSP 2006, vol. 1, May 2006, pp. I–I.

[7] Online: http://www.dsp.sun.ac.za/~nbrummer/focal/