The d-Ear and CST systems for NIST 2006 SRE

Thomas Fang Zheng

Beijing d-Ear Technologies Co., Ltd.
http://www.d-Ear.com
and
Center for Speech Technology,
National Lab for Information Science and Technology,
Tsinghua University, Beijing, 100084, China
http://cst.cs.tsinghua.edu.cn/

Outline

- Introduction
- System description
- Results
- Future work

Introduction

- Beijing d-Ear Technologies Co., Ltd.
  - Speaker recognition group
- Center for Speech Technology (CST), Tsinghua National Lab for Information Science and Technology, Tsinghua University
  - Speaker recognition group

- Involved tasks
  - 1conv4w train - 1conv4w test
  - 1conv4w train - 10sec4w test
  - 1conv4w train - 1conv2w test

System Description

- Overview
  - MFCC features
  - GMM-UBM structure
  - T-Norm score normalization
  - Speech segmentation and clustering for 2-speaker conversations

Feature Extraction

- 16-dimemsional MFCC plus delta
- Bandwidth: 100 ~ 3800 Hz
- Hamming window with 20ms' length and 10ms' shift
- Pitch-based silence elimination / Energy-based silence elimination
- CMS and CVN

GMM-UBM Systems

- Two gender-dependent UBMs
- 1,024 Gaussian components in each UBM
- Trained with channel-balanced (landline, cellular, cordless) speech from NIST 2004 SRE

Speaker Model Adaptation

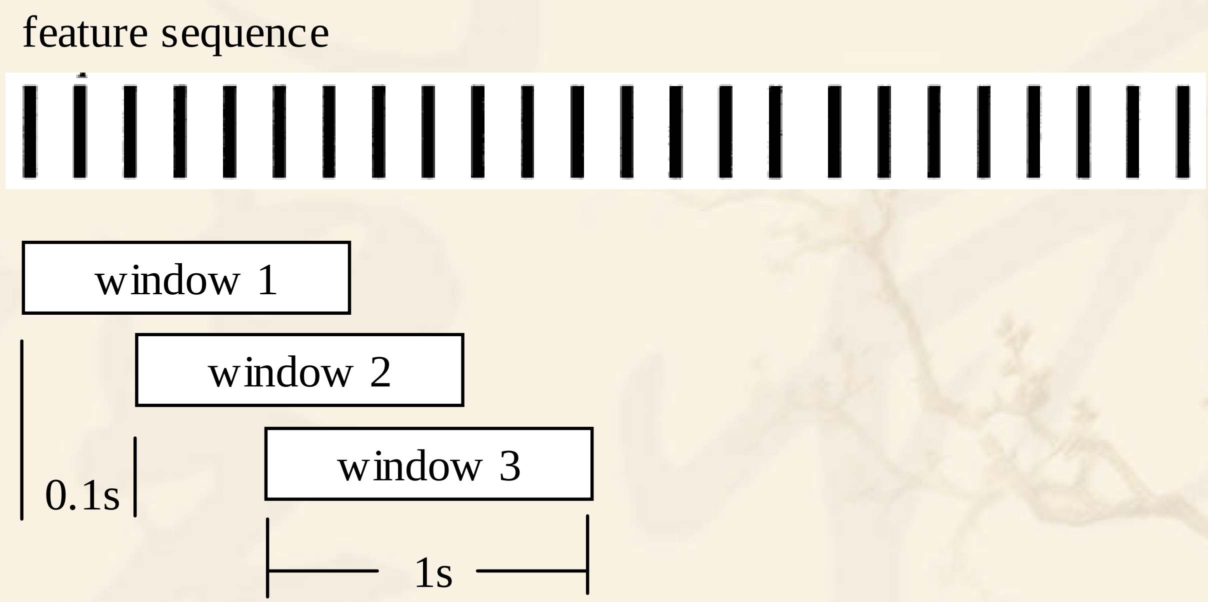
- MAP adaptation
  - Relevance factor automatically adjusted
  - Only means adapted

Score Normalization

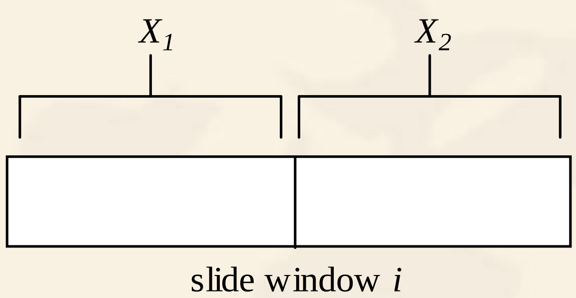
- T-Norm
  - 368 female speakers
  - 248 male speakers
  - Selected from NIST 2004 SRE for the calculation of T-Norm parameters

Speech Segmentation based on Log Likelihood Ratio Score (LLRS) over UBM

- A slide window (with 2s' length and 0.1s' shift) is applied on the conversational speech



- The feature sequence in each window is divided into 2 parts (X_1, X_2)
- These two parts are scored against a 1,024-component UBM, and their log likelihood ratio score (LLRS) are computed.



$$\Delta S(i) = abs\left(L\left(X_1 \mid UBM\right) - L\left(X_2 \mid UBM\right)\right)$$

- For each conversation, a sequence of LLRS' can be obtained, and their standard deviation σ is estimated.
- In the LLRS plot, a peak is assumed to be a speaker change point

$$|max\text{-}min_l| > \alpha\sigma \quad \text{and} \quad |max\text{-}min_r| > \alpha\sigma$$

where *max* is the LLRS of a peak, min_l and min_r are the left and right minima next to the peak, and α is an experiential value which is set to 0.5

Speaker Clustering

- **Initialization**

  Step 1.1: an initial speaker model S_0 is adapted on the whole conversation from a 16-component UBM;

  Step 1.2: each speech segment is scored against S_0, and the segment longer than 2s and with the maximum score is selected to adapt speaker model S_1 from the 16-component UBM;

  Step 1.3: The remaining segments are scored against S_0 and S_1, respectively. The score difference ΔS is computed as $\Delta S = L(X|S_0) - L(X|S_1)$.

  The segments longer than 2s and with the maximum ΔS is selected to adapt speaker model S_2 from the 16-component UBM.

- Iterations:

Step 2.1: Score the remaining segments against speaker model S_1 and S_2, ΔS_{12} and ΔS_{21} are computed,

$$\Delta S_{12} = L(X|S_1) - L(X|S_2) ,$$

$$\Delta S_{21} = L(X|S_2) - L(X|S_1) ,$$

The segment longer than 1s and with maximum ΔS_{12} is assigned to S_1 and used to update S_1; the segment longer than 1s and with maximum ΔS_{21} is assigned to S_2 and used to update S_2;

Step 2.2: Repeat step 1 until there is no speech longer than 1s;

- **Refinement**

  Step 3.1: all the segments in the conversation are scored against speaker model S_1 and S_2, and corresponding ΔS_{12} and ΔS_{21} are computed;

  Step 3.2: use segments whose ΔS_{12} is among the top half of all the positive ΔS_{12} to adapt a new speaker model S_1' from a 1,024-component UBM;

  Step 3.3: use segments whose ΔS_{21} is among the top half of all the positive ΔS_{21} to adapt a new speaker model S_2' from a 1,024-component UBM;

  Step 3.4: use S_1' and S_2' to reclassify all the segments in the conversation into 2 clusters.

- Segmentation criterion:
  - GLR for d-Ear system
  - UBM LLRS based segmentation for CST system

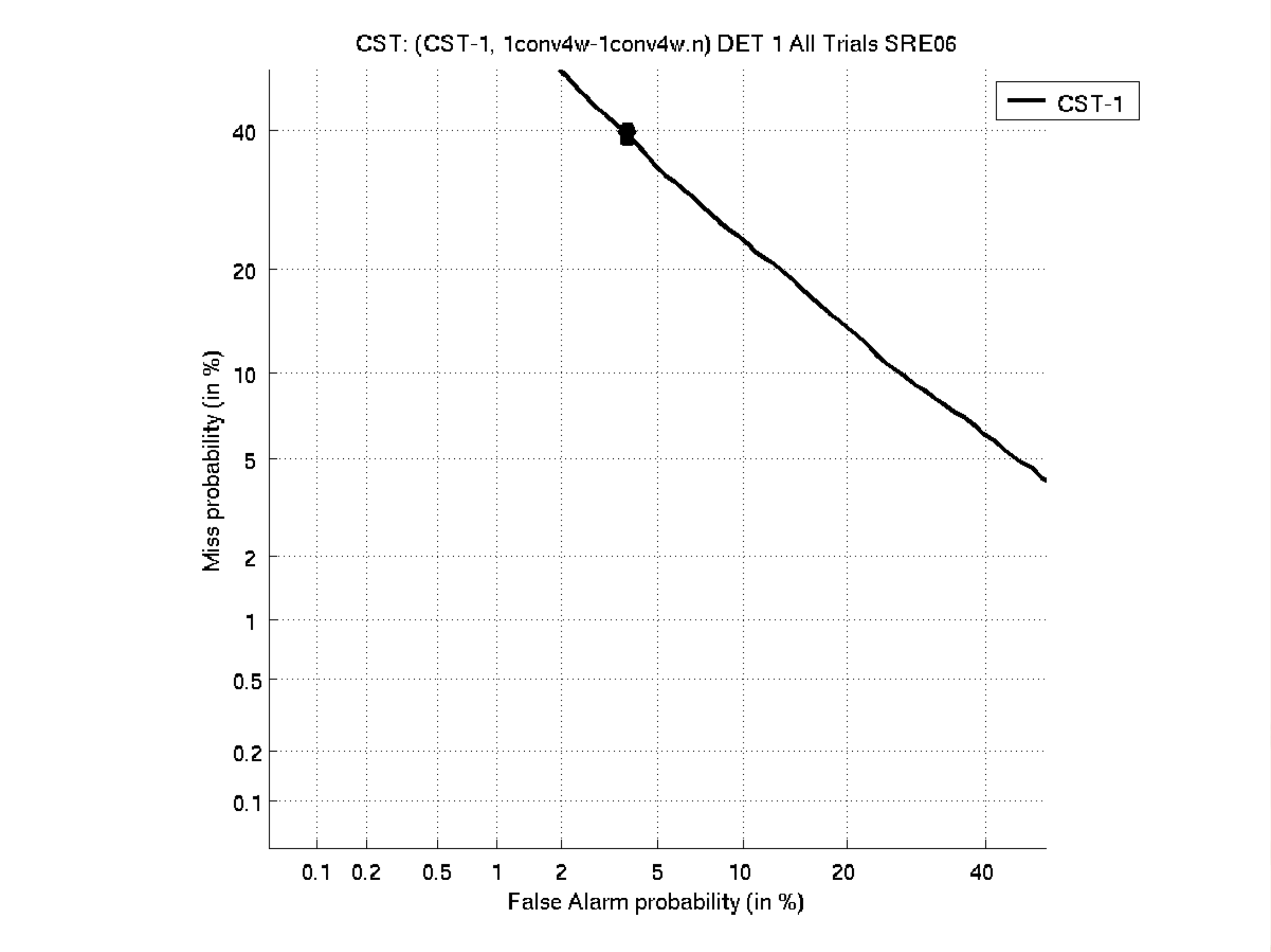
Results of Speaker Segmentation and Clustering

- Results on NIST2002 switch board conversation segmentation tasks

Error Type	Error Time Rate (%)
Missed Speaker Time	0.1
False Alarm Speaker Time	0.1
Speaker Error Time	6.6

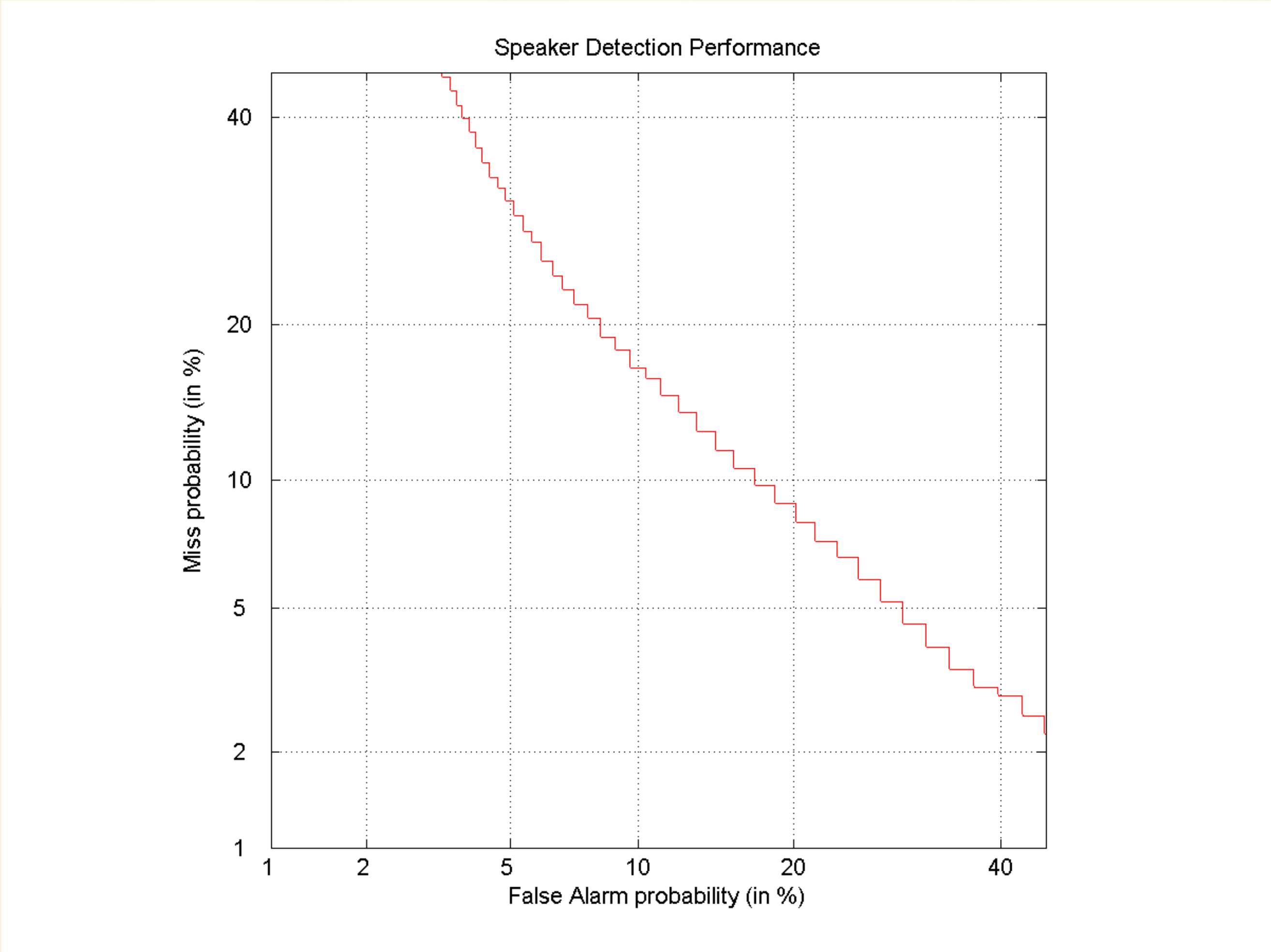
Results on NIST 2006 SRE

- 1c4w-1c4w (CST)
  - Pitch-based silence elimination -- submitted results

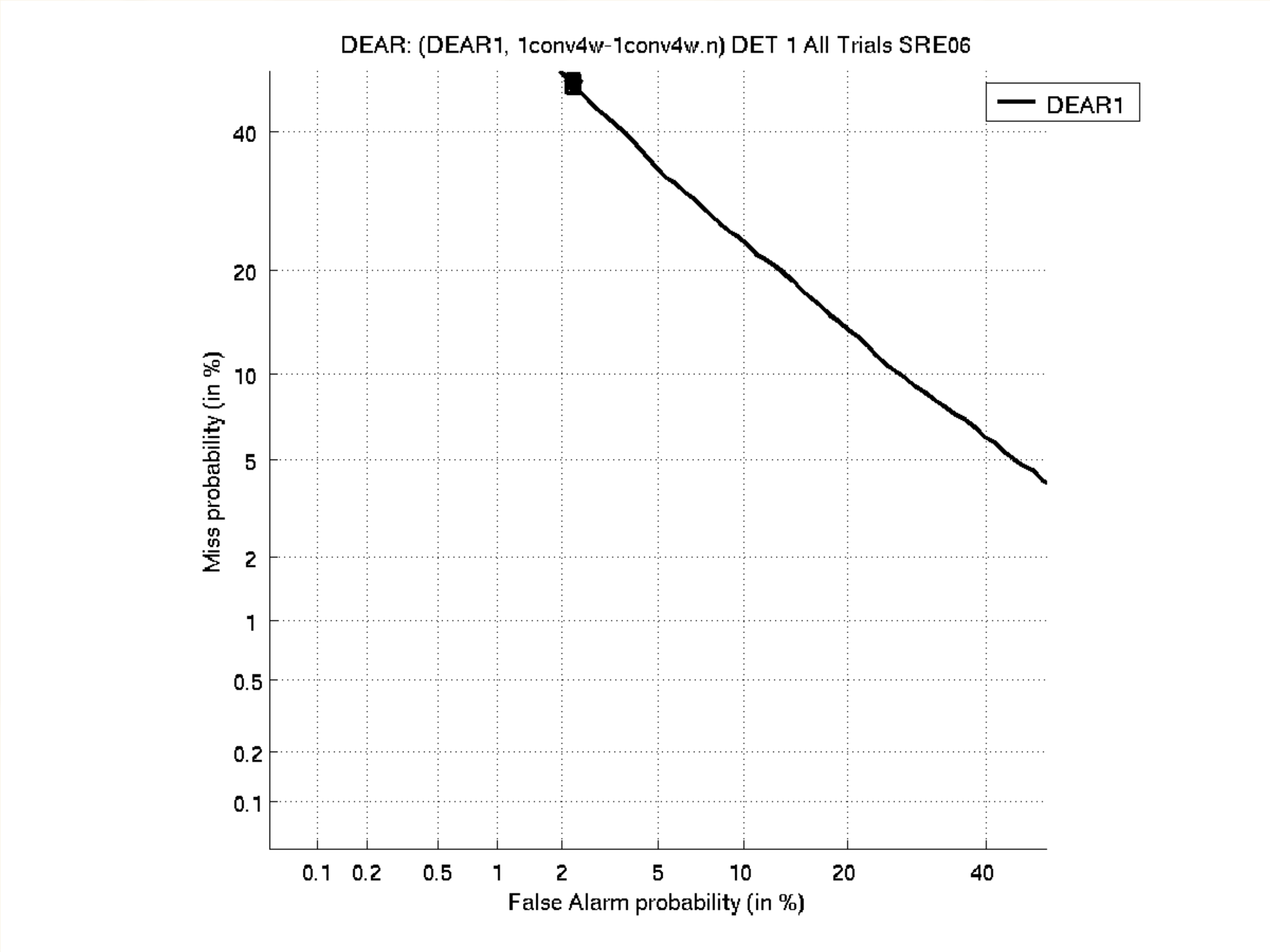


v 1c4w-1c4w (CST)

***Energy-based** silence elimination -- new results*

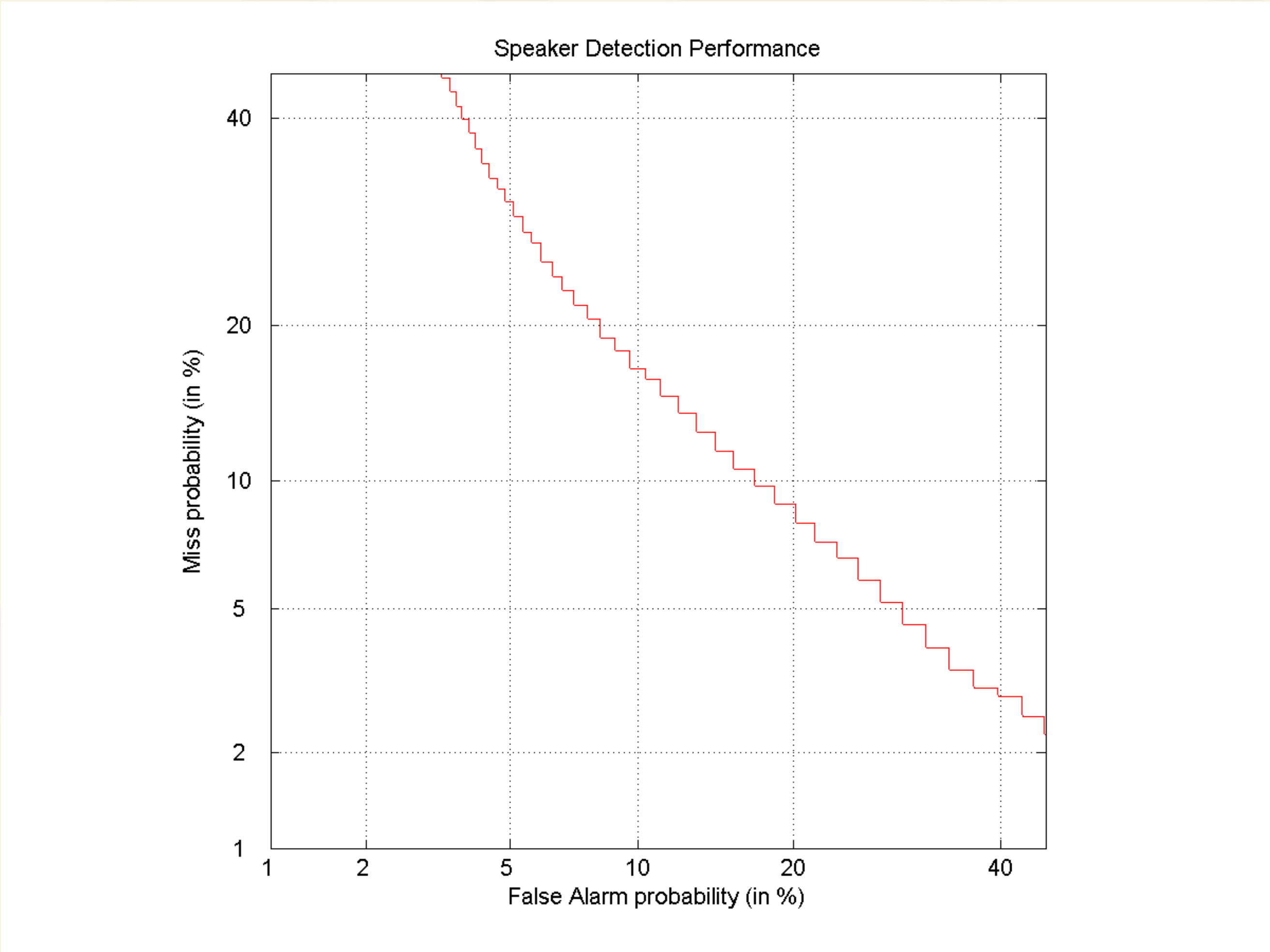


- 1c4w-1c4w (d-Ear)
  - Pitch-based silence elimination -- submitted results



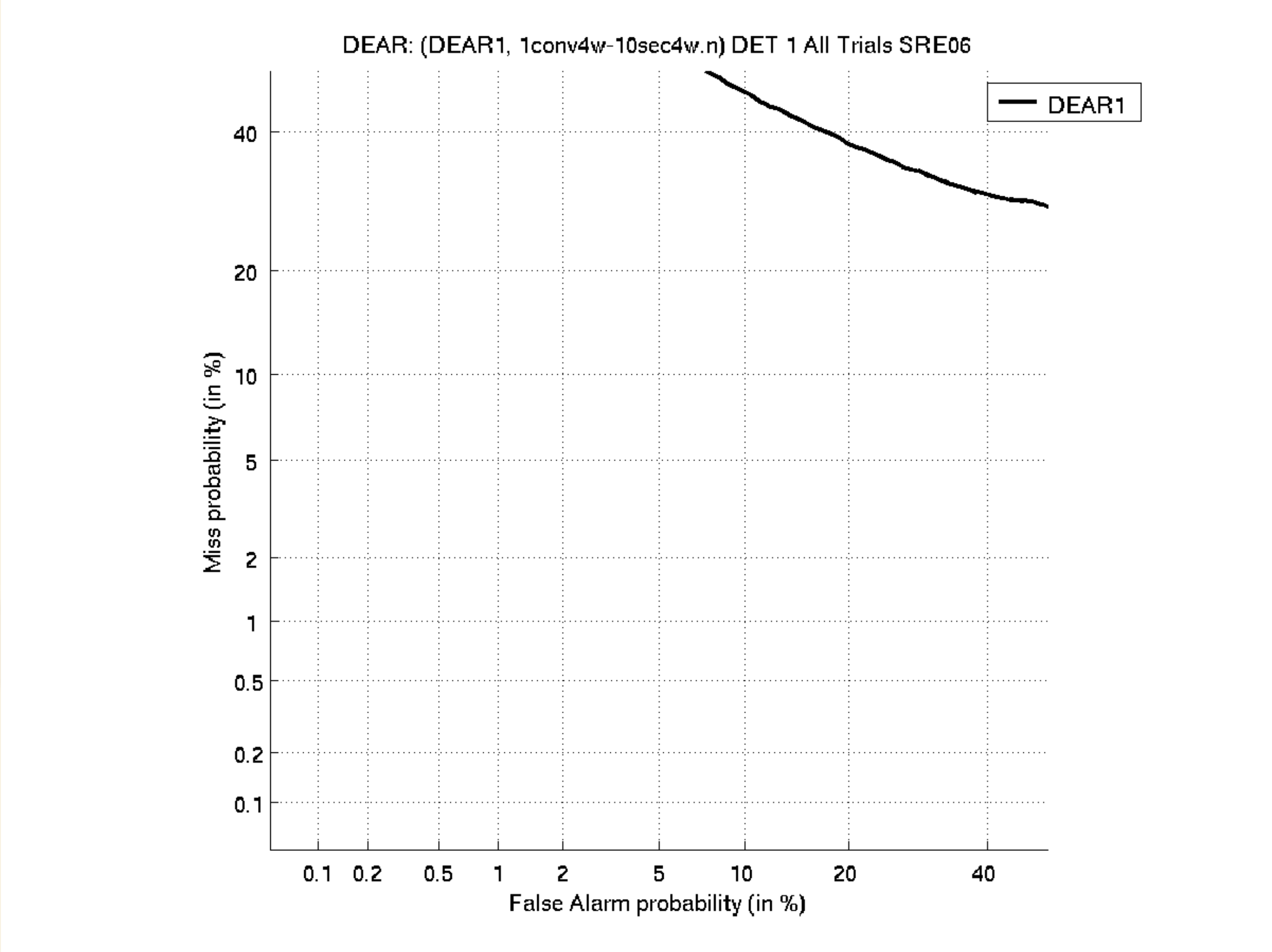
v 1c4w-1c4w (d-Ear)

***Energy-based** silence elimination -- new results*



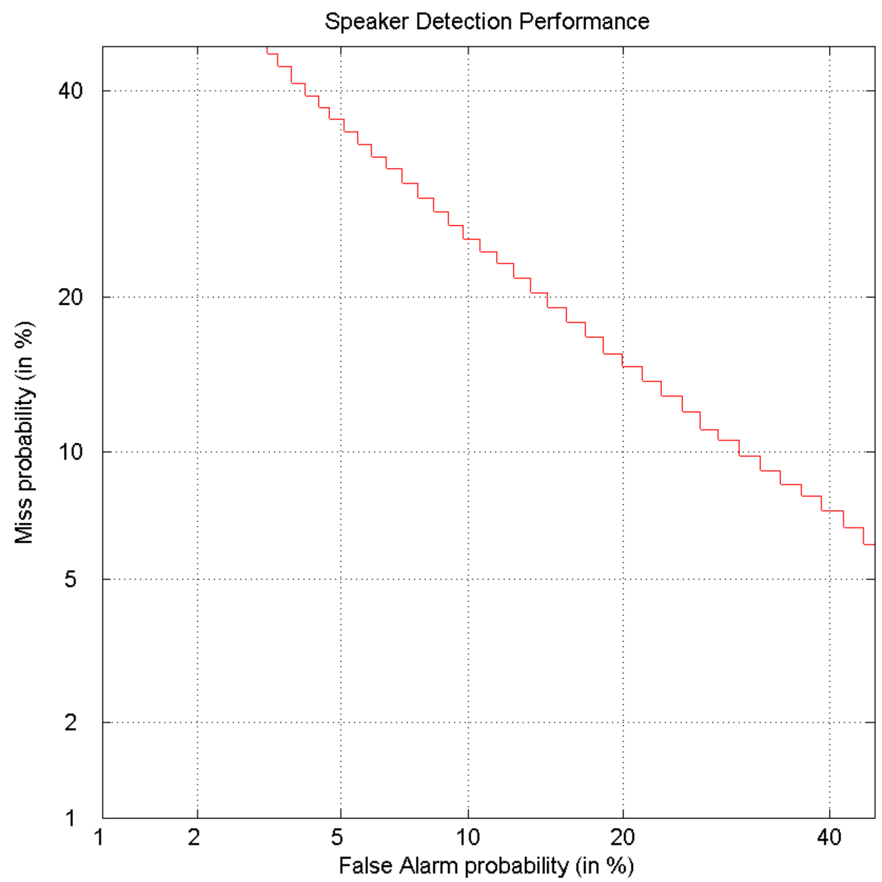
1c4w-10sec4w (d-Ear)

pitch-based silence elimination -- submitted results

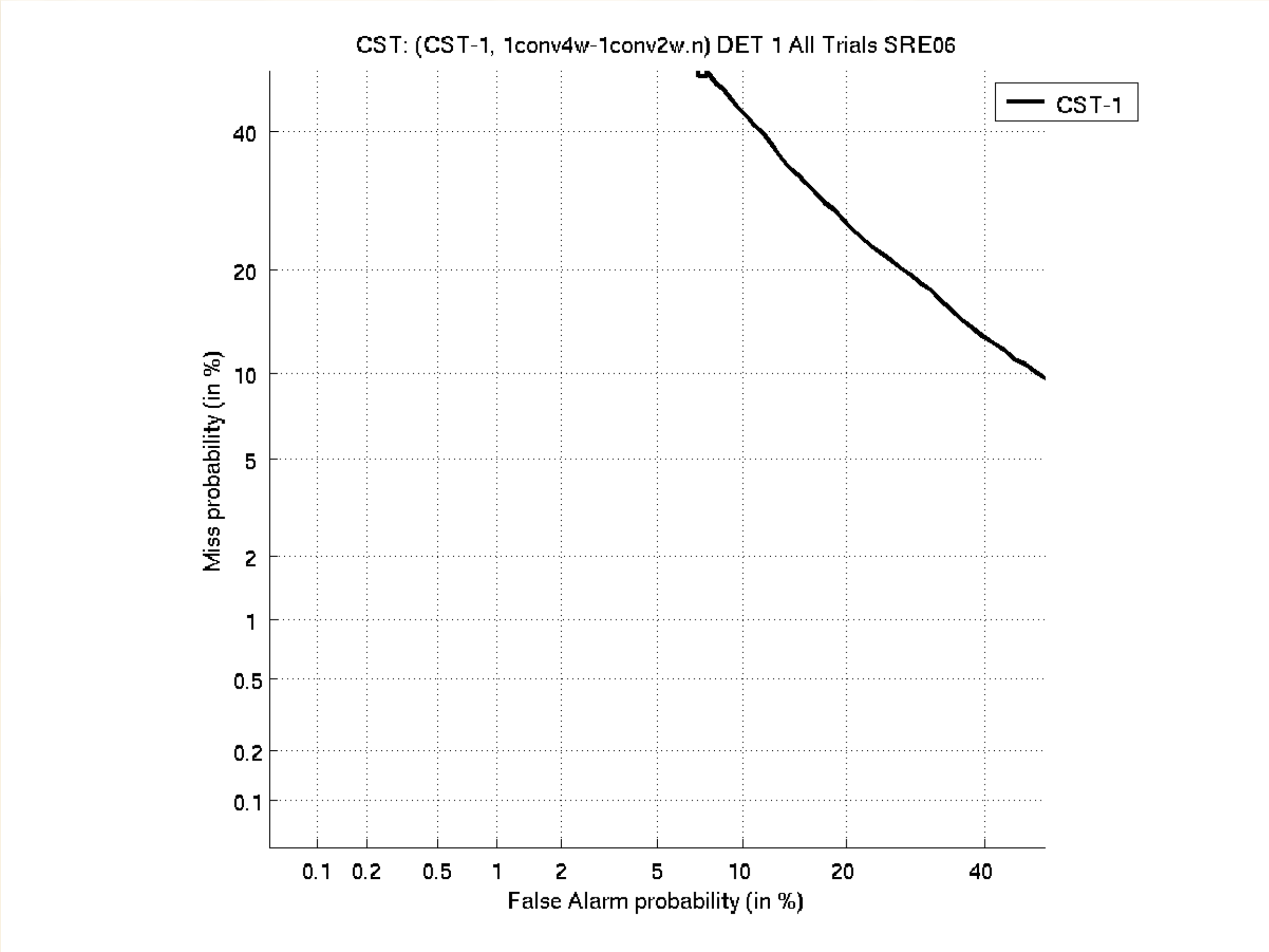


- 1c4w-10sec4w (d-Ear)

  ***Energy-based** silence elimination -- new results*

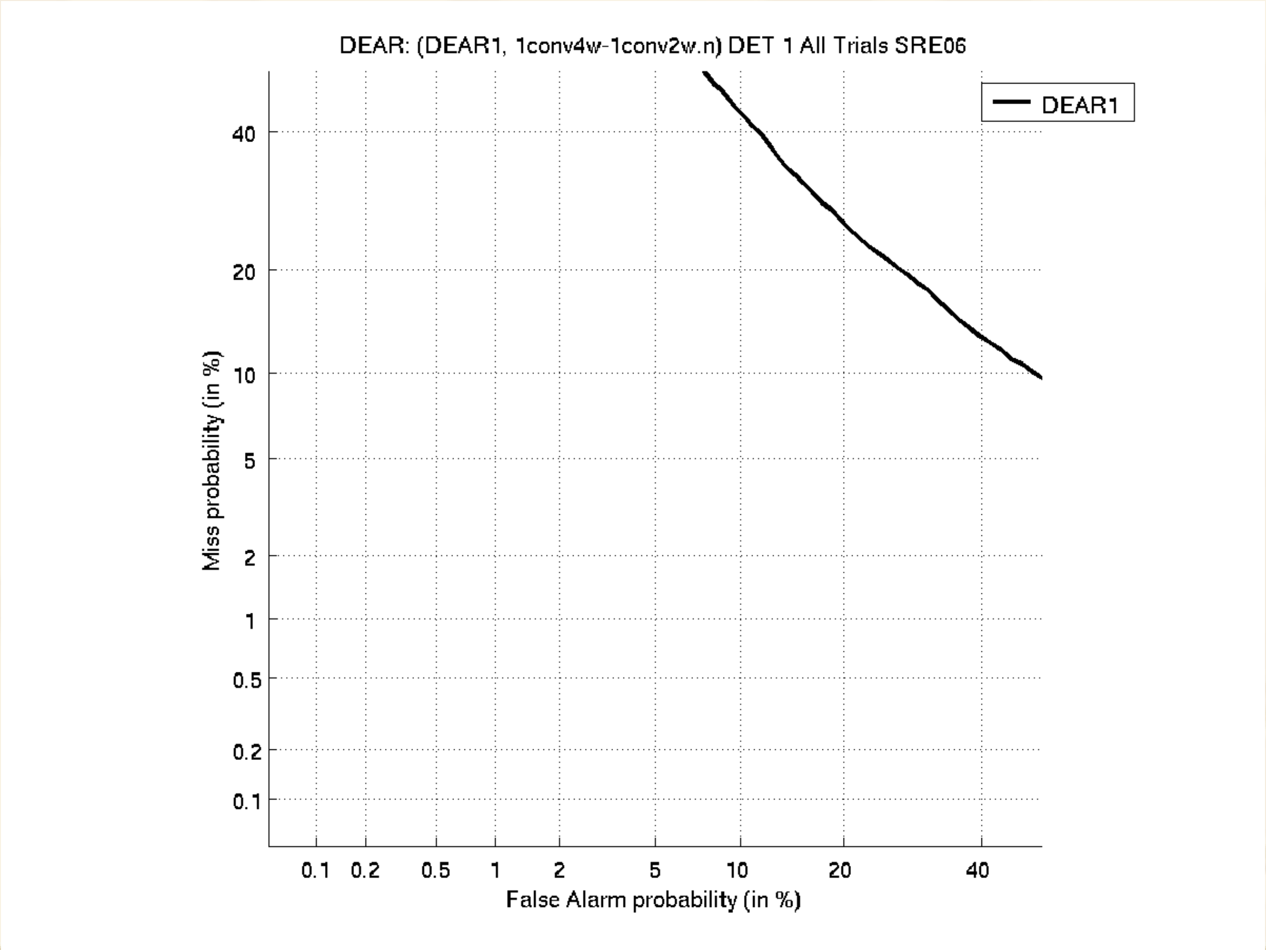


- 1c4w-1c2w (CST)
  - Pitch-based silence elimination -- submitted results



1c4w-1c2w (CST)

Pitch-based silence elimination -- submitted results



Remarks

- Pitch-based silence elimination
  - Using pitch information for VAD, which is better for application in noisy environments
  - yet reserving shorter speech segments

- ***Energy-based*** *silence elimination*
  - *Using frame energy information for VAD, reserving longer speech segments*
  - *Better in relatively cleaner environments*

Thank You !