



CSD-UOC

ROBUST TEXT-INDEPENDENT SPEAKER IDENTIFICATION USING SHORT TEST AND TRAINING SESSIONS

Christos Tzagkarakis and Athanasios Mouchtaris

Institute of Computer Science - Foundation for Research & Technology-Hellas



FORTH-ICS

1

INTRODUCTION

- > Speaker recognition systems
 - security
 - commercial applications
- > focus ---> recognition using only voice patterns
- > Categories according to matching
 - > **verification**: verify speaker's claim that is of a certain identity
 - > **identification**: determine an unknown speaker's identity
- > Categories according to training process
 - > **text-dependent**: same text being used for both training and recognition
 - > **text-independent**: final recognition result does not rely on a specific text

Goal of the present study ---> noise-robust text independent speaker identification (SID) under the assumption of short testing & training conditions

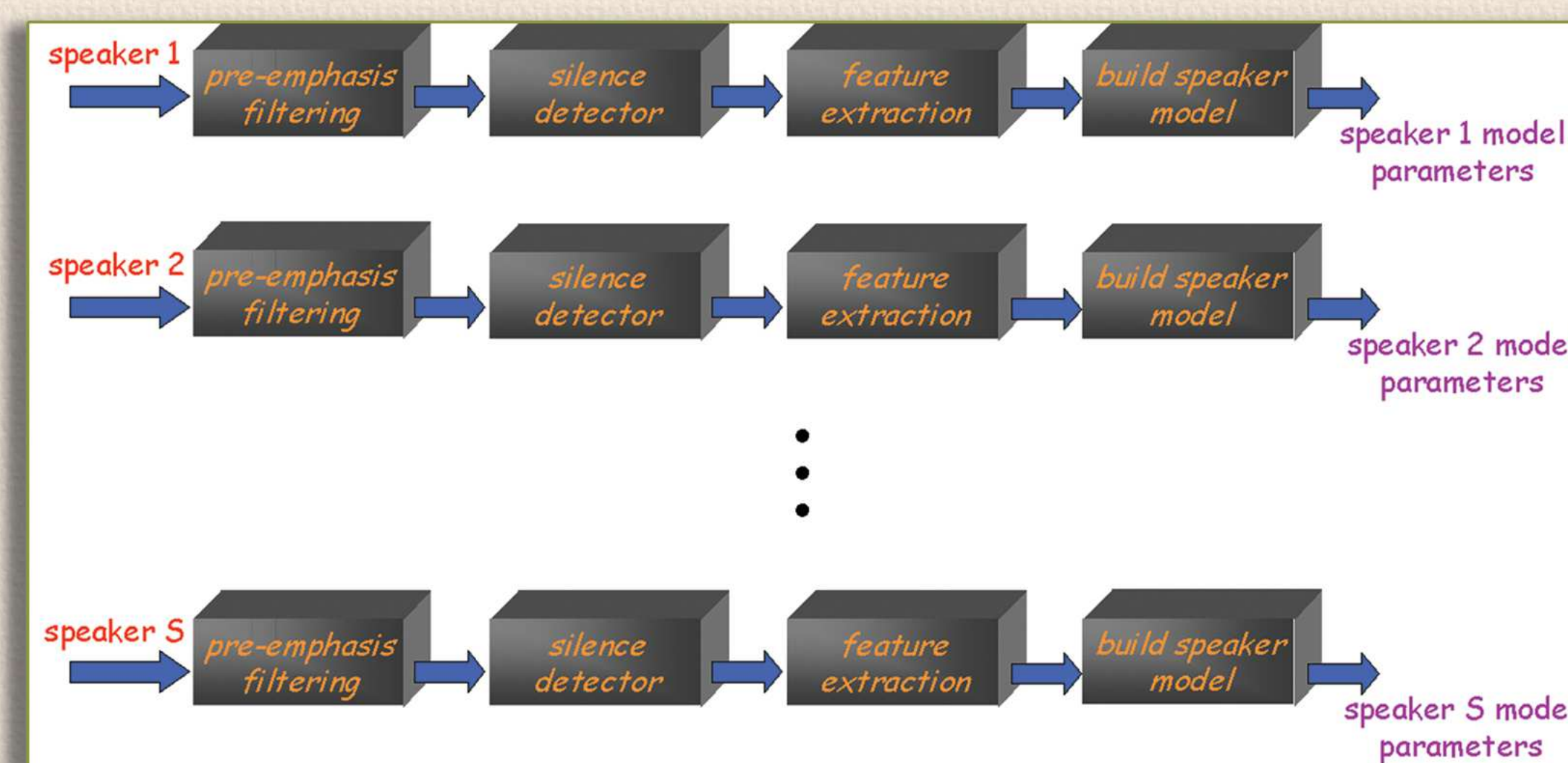
Why to follow this approach?

- > speech signals are corrupted by noise due to either the environment where the speaker is present or due to the voice transmission medium
- > it is often not feasible to have large amounts of training data from all the speakers
- > to speed up the SID process the speaker utterance to be identified (testing data) should be as short as possible

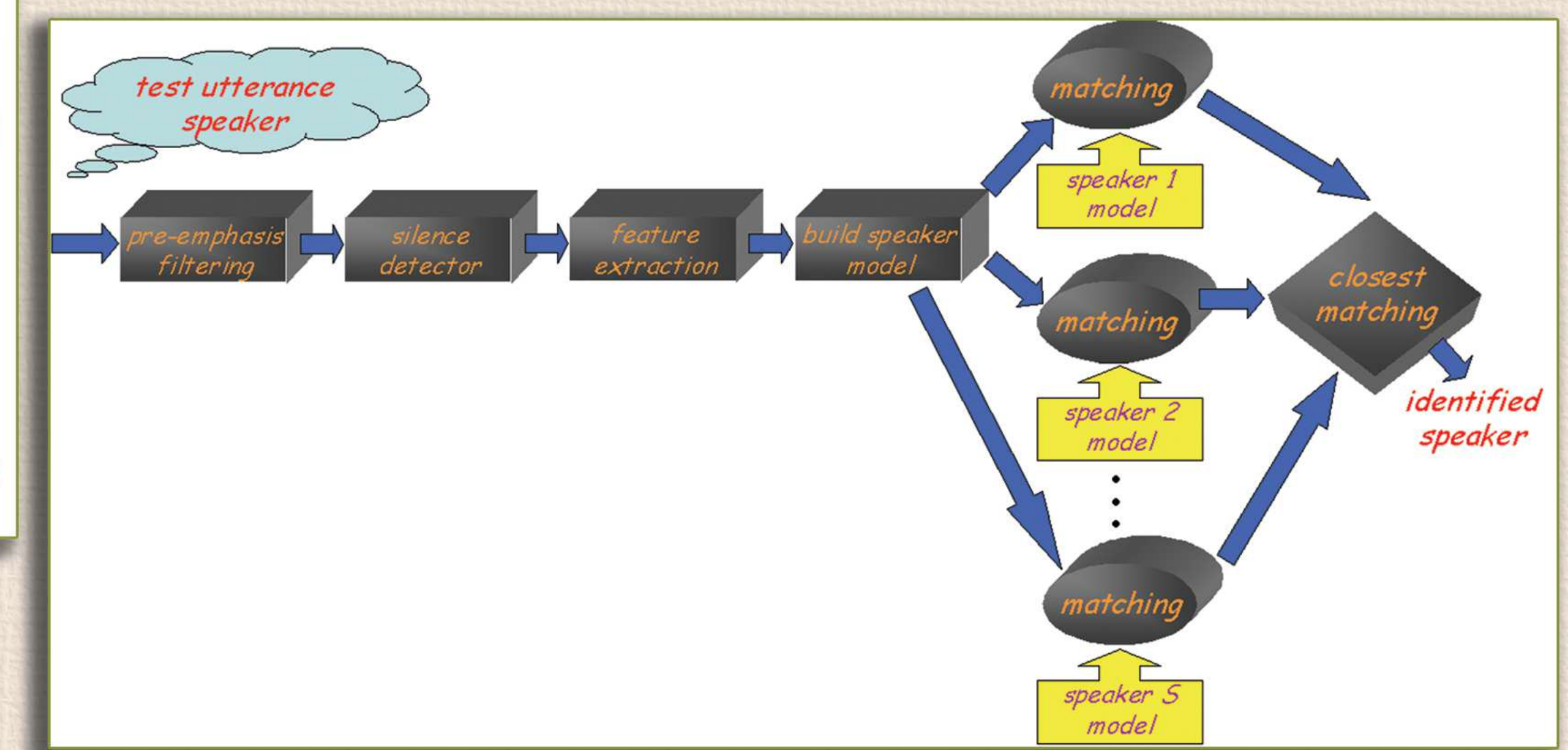
2

ARCHITECTURE OF AN SID SYSTEM

I. TRAINING PHASE



II. TESTING PHASE



3

CLASSIFICATION METHODS

A Gaussian Mixture Model

- > model the PDF of the feature space of each speaker in the dataset (training phase) as a sum of Gaussian functions
- > apply maximum a-posteriori rule to identify the speaker
- > Gaussian mixture density ---> weighted sum of M multi-dimensional Gaussian densities

$$\lambda_i = \left\{ p_m^i, \mu_m^i, \Sigma_m^i \right\}, m = 1, \dots, M$$

i^{th} speaker's prior probability of the m^{th} mixture

- > represent each speaker by Gaussian Mixture Model (GMM) and the corresponding model λ
 - > parameters are computed via Expectation-Maximization
- > for the identification task (testing phase) ---> estimated speaker identity (speaker index) obtained based on MAP rule for a given sequence of observations

$$S_q = \arg \max_{1 \leq i \leq S} p(\lambda_i | X) = \arg \max_{1 \leq i \leq S} \frac{p(X | \lambda_i) p(\lambda_i)}{p(X)}$$

$$S_q = \arg \max_{1 \leq i \leq S} p(X | \lambda_i) \quad \text{assume independent observations (take the logarithm)}$$

$$p(x_t | \lambda_i) = \prod_{m=1}^M \frac{p_m^i}{(2\pi)^{K/2} |\Sigma_m^i|^{1/2}} \exp \left\{ -\frac{1}{2} (x_t - \mu_m^i)^T \Sigma_m^i^{-1} (x_t - \mu_m^i) \right\}$$

B Kullback-Leibler Divergence using Generalized Gaussian Density

- > treat SID task as a multiple hypothesis problem
 - > given a set X of feature vectors $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,K})^T$ ---> each speaker is assigned a hypothesis H_i
 - > select one hypothesis out of S which best describes the data from S_q

$$p(x_t | H_j) \geq p(x_t | H_i), i \neq j, \forall i = 1, \dots, S$$

- > each conditional probability $p(x | H)$ modeled by $p(x; \theta)$ ---> θ_i set of model parameters

- > compute the Kullback-Leibler Divergence (KLD)

$$D(p(x; \theta_q) \| p(x; \theta_i)) = \int p(x; \theta_q) \log \frac{p(x; \theta_q)}{p(x; \theta_i)} dx$$

- > smallest value of KLD ---> identified speaker S_q
- > chain rule holds for the KLD ---> applied in order to combine the KLDs from multiple datasets or dataset dimensions
 - > fit a Generalized Gaussian Density (GGD) on the PDF of the features

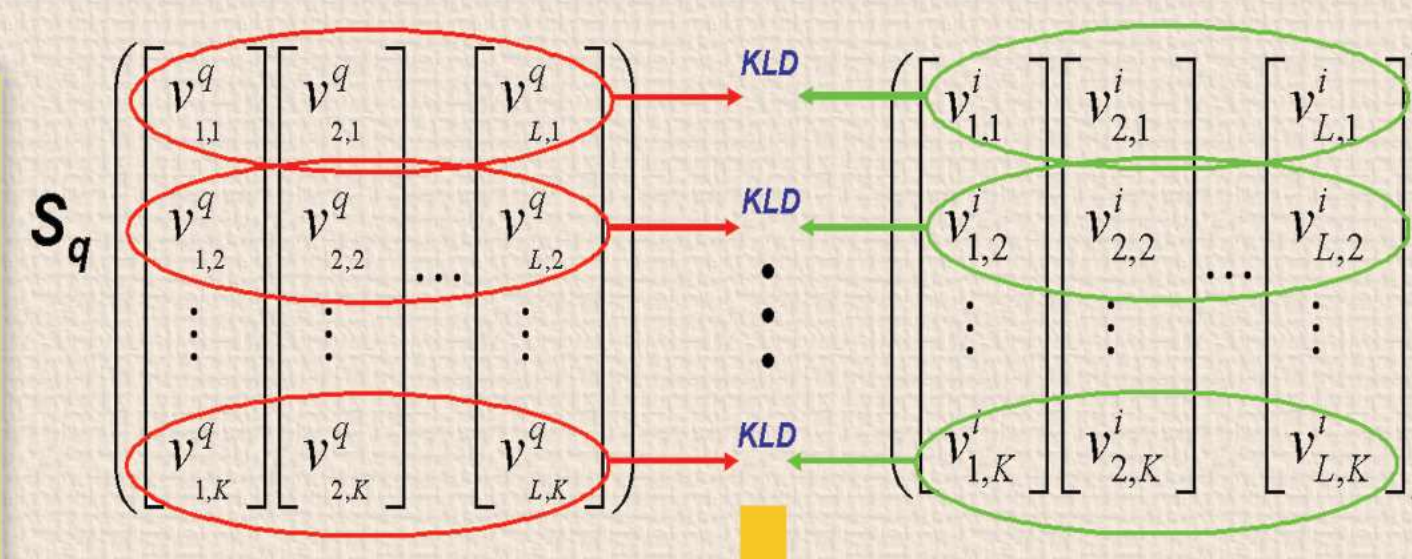
$$D(p(X, Y) \| q(X, Y)) = D(p(X) \| q(X)) + D(p(Y) \| q(Y))$$

Generalized Gaussian Density

$$p(x; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x|/\alpha)^\beta}$$

KLD between two GGDs

$$D(p_{\alpha_1, \beta_1} \| p_{\alpha_2, \beta_2}) = \log \left(\frac{\beta_1 \alpha_2 \Gamma(1/\beta_2)}{\beta_2 \alpha_1 \Gamma(1/\beta_1)} \right) + \left(\frac{\alpha_1}{\alpha_2} \right)^\beta \frac{\Gamma(\frac{\beta_2+1}{\beta_1})}{\Gamma(\frac{\beta_1}{\beta_1})} - \frac{1}{\beta_1}$$



$$D(S_q \| S_i) = \frac{1}{K} \sum_{k=1}^K D(p_{S_q,k} \| p_{S_i,k})$$

C Sparse Representation Classification

- > assume S speakers (classes)
 - > focus on the i^{th} speaker
 - > n_i training samples (feature vectors)

$$V_i = [v_{i,1} | v_{i,2} | \dots | v_{i,n_i}] \in \mathbb{R}^{K \times n_i}$$

- > given a new test sample (feature vector) ---> $x_t \in \mathbb{R}^K$

$$x_t = c_{i,1} v_{i,1} + c_{i,2} v_{i,2} + \dots + c_{i,n_i} v_{i,n_i} = V_i c_i$$

- > entire training set ---> concatenation of the $N = n_1 + \dots + n_S$ training samples of all S speakers

$$V = [V_1 | V_2 | \dots | V_S] = [v_{1,1} | v_{1,2} | \dots | v_{1,n_1} | \dots | v_{S,1} | \dots | v_{S,n_S}]$$

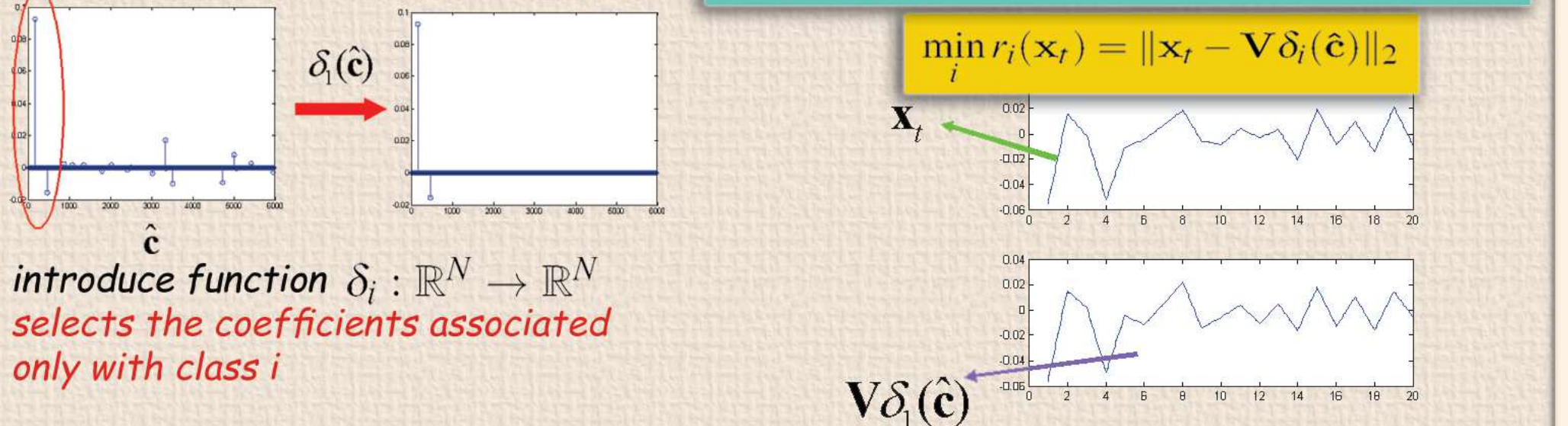
$$c = [0, \dots, 0, c_{i,1}, c_{i,2}, \dots, c_{i,n_i}, 0, \dots, 0]^T \in \mathbb{R}^N$$

- > **motivation** ---> seek the sparsest solution to $x_t = Vc$

$$\hat{c} = \arg \min_c \|c\|_1, \text{ s.t. } x_t = Vc$$

- > approximate the test sample x_t as $V\delta(\hat{c})$

$$\min_r r_t(x_t) = \|x_t - V\delta(\hat{c})\|_2$$



EXPERIMENTAL RESULTS

4

- > VOICES corpus ---> 7 male & 5 female speakers
 - > sampling rate 22050 Hz

- > pre-emphasis filtering ---> low-pass $H(z) = 1 - 0.97z^{-1}$
- > silence detector ---> short-term energy & zero-crossings
- > feature extraction ---> mel-frequency cepstral coefficients order 20

- > analysis window 20 msec (50% overlapping)

Performance evaluation

$$\underbrace{x_1, x_2, x_3, \dots, x_L, x_{L+1}, x_{L+2}, \dots}_{1^{\text{st}} \text{ segment}}$$

Correct identification rate

$$\text{correct ident. rate} = \frac{\# \text{ correctly identified segments}}{\text{total \# of segments}} \cdot 100\%$$

- > noise robust SID ---> AWGN @ 10,15,20,25 dB
- > segment length ---> L=10:40:500
- > training utterances (clean speech data) duration ---> 5,10,15,20 sec

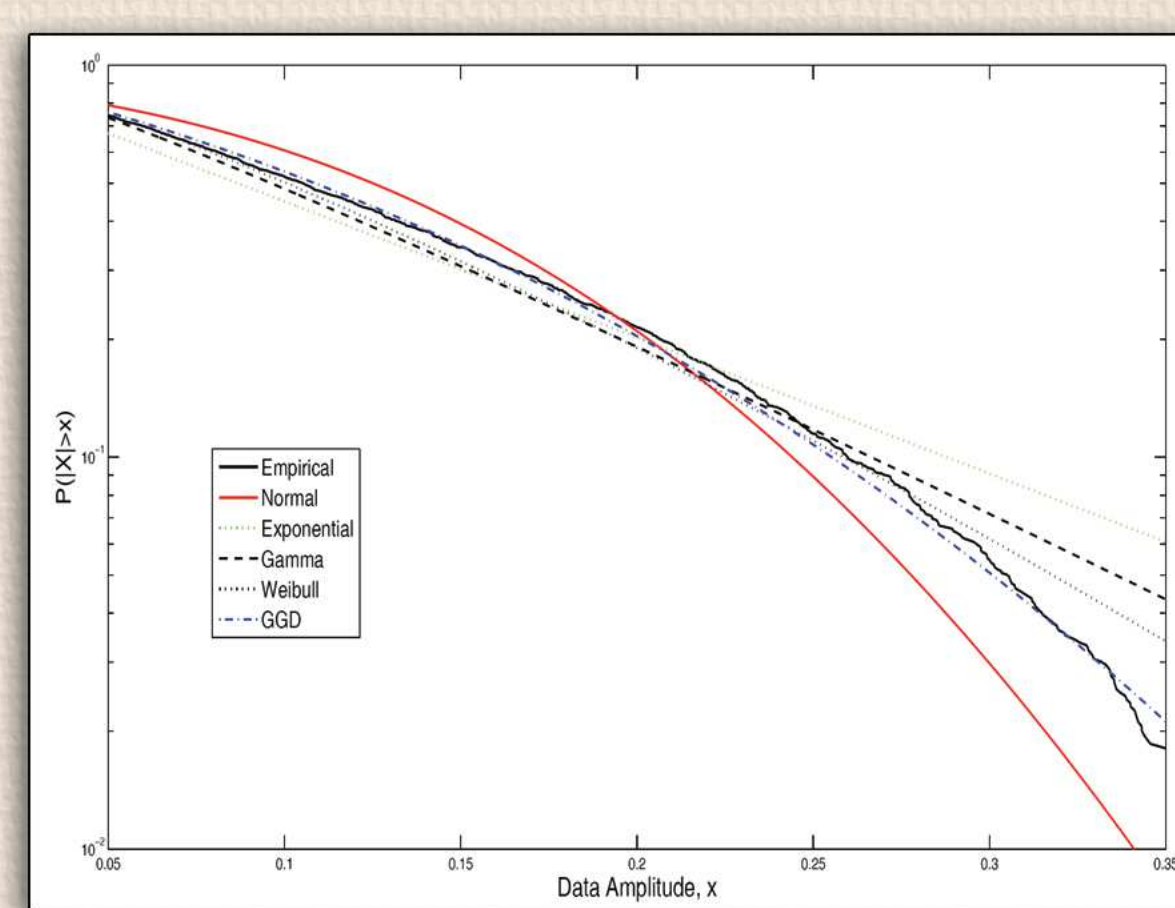


Fig. 1: Example Amplitude Probability Density curves of the 8th MFCC coefficient from the training data (20 sec) of the 10th speaker.

(1) SID using GMM

- # of mixtures 4 ---> 5 & 10 sec training data
- # of mixtures 8 ---> 15 & 20 sec training data

(2) SID using KLD based on GGD

- # of mixtures 4 ---> 5 & 10 sec training data

(3) SID using SRC

- n_i # frames per speaker ---> size of V : $20 \times (12 \times n_i)$

$$\hat{c}_l = \arg \min_{c_l} \|c_l\|_1, \text{ s.t. } x_{t,l} = Vc_l, \text{ for } l = 1, \dots, L$$

- 5 sec training data dur., utterance length ~2 sec
 - correct SID perf. > 80% for 15, 20 & 25 dB SNR
- 5 sec training data dur., utterance length ~2 sec, 10 dB SNR
 - 70% correct SID perf.
- 20 sec training data dur., 1.5 sec test data
 - correct SID perf. > 70% for all SNR values

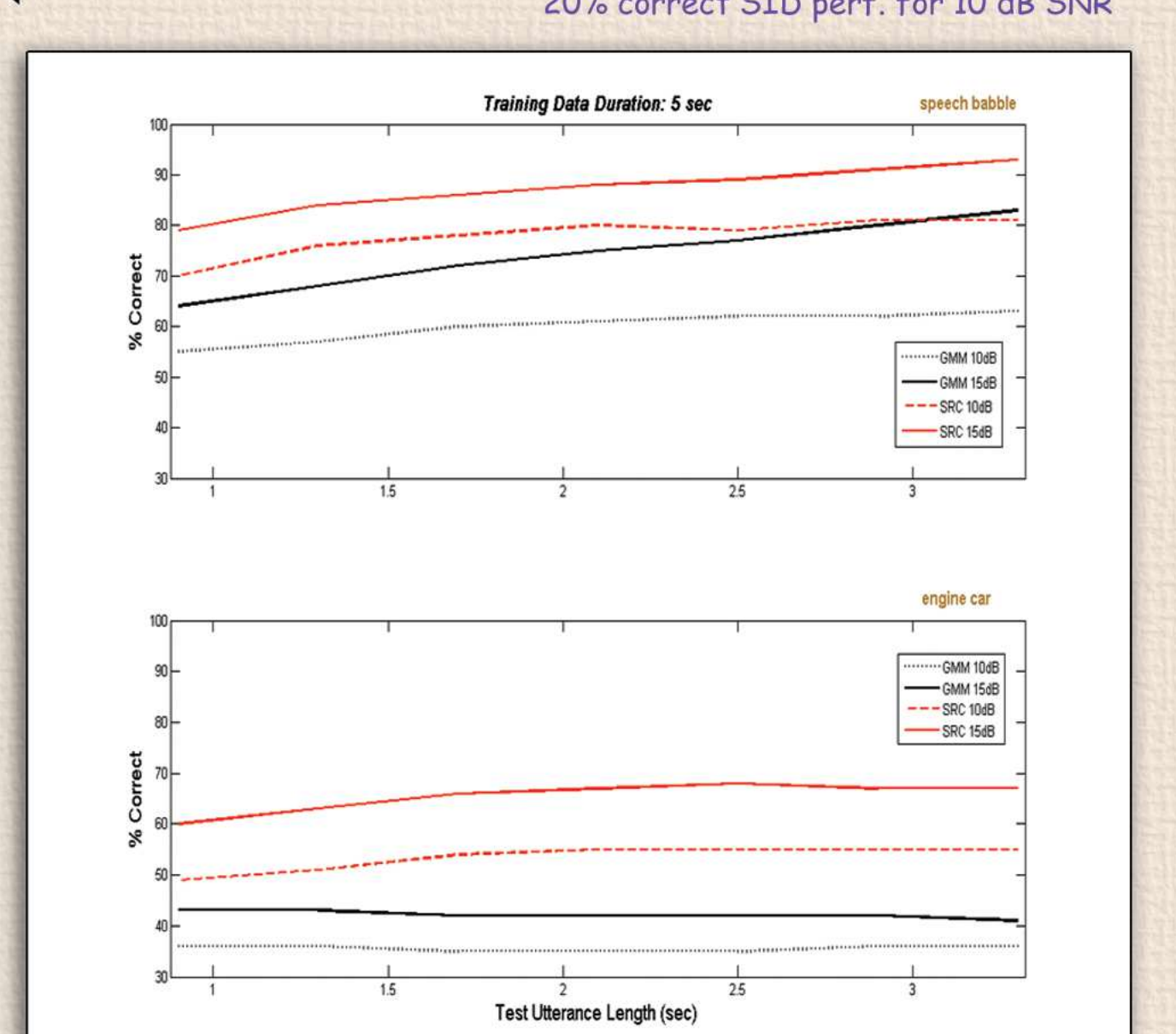


Fig. 3: SID performance as a function of the test data duration for speech-babble and engine-car noise (10dB & 15dB).