SIPL SIGNAL & IMAGE PROCESSING LAB Technion Israel Institute of Technology

Speech Enhancement for Speech Recognition using Particle Filtering

Students:

Asa Dan & Elad Shimoni

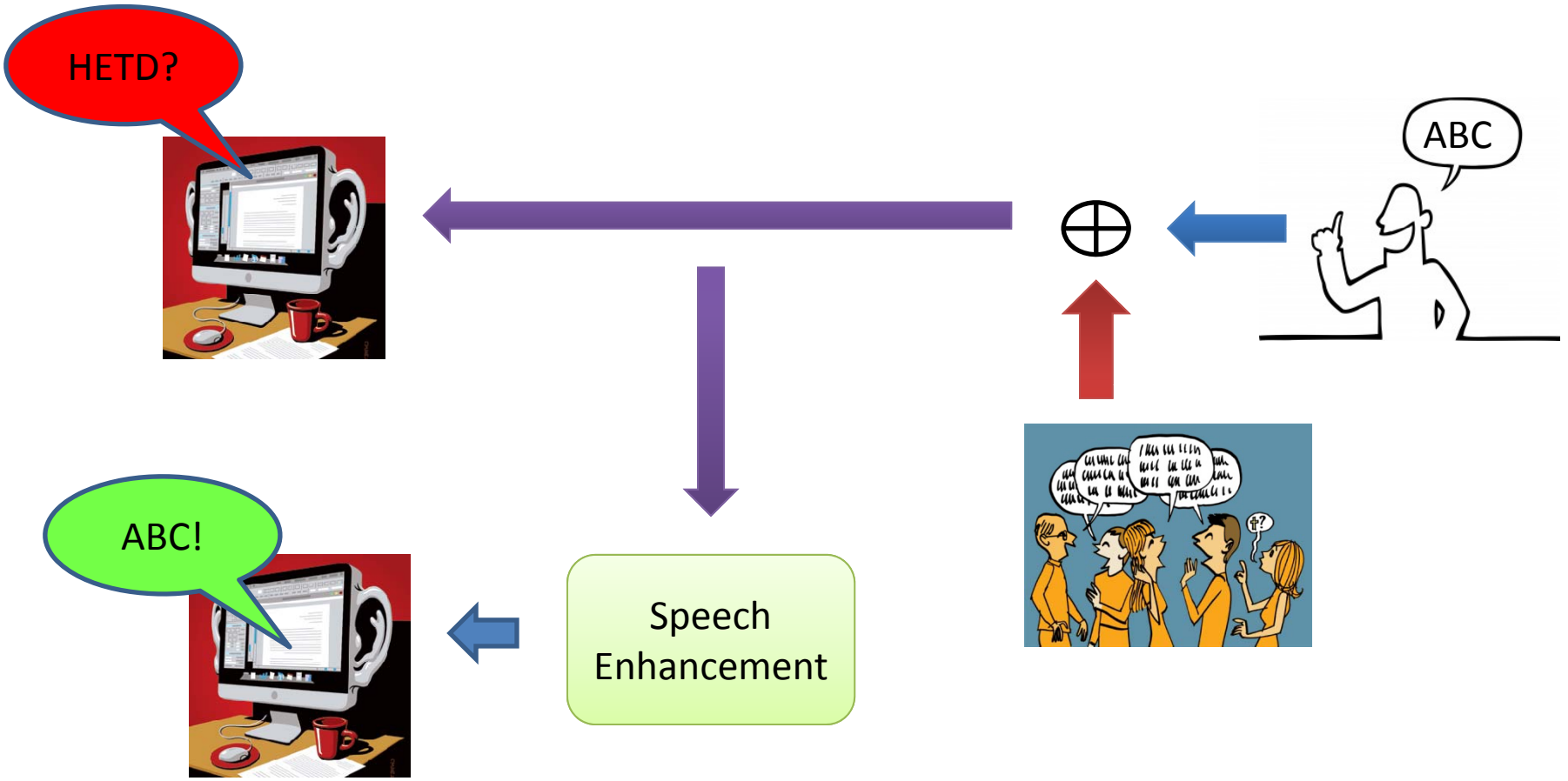
Instructor:

Hadas Benisty

Motivation

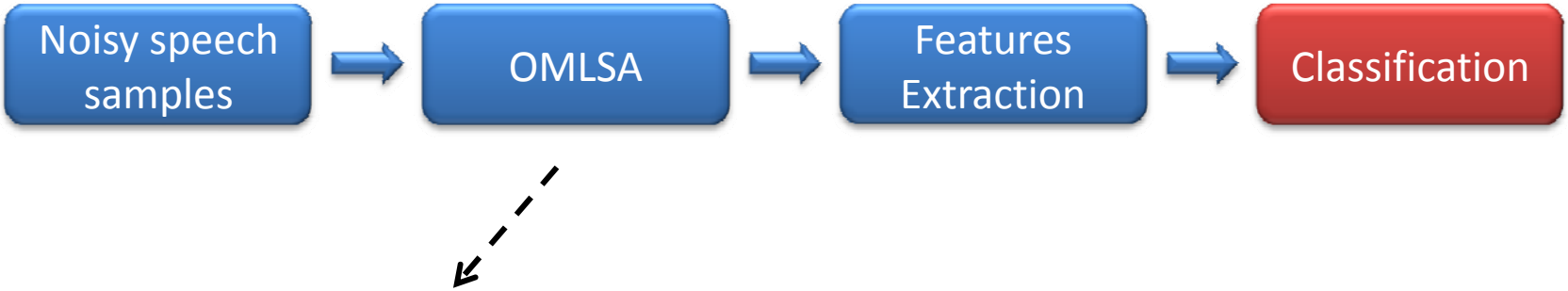
Motivation

Improved **Speech Recognition** in noisy environment



Proposed Solution #1

A previous project[1]

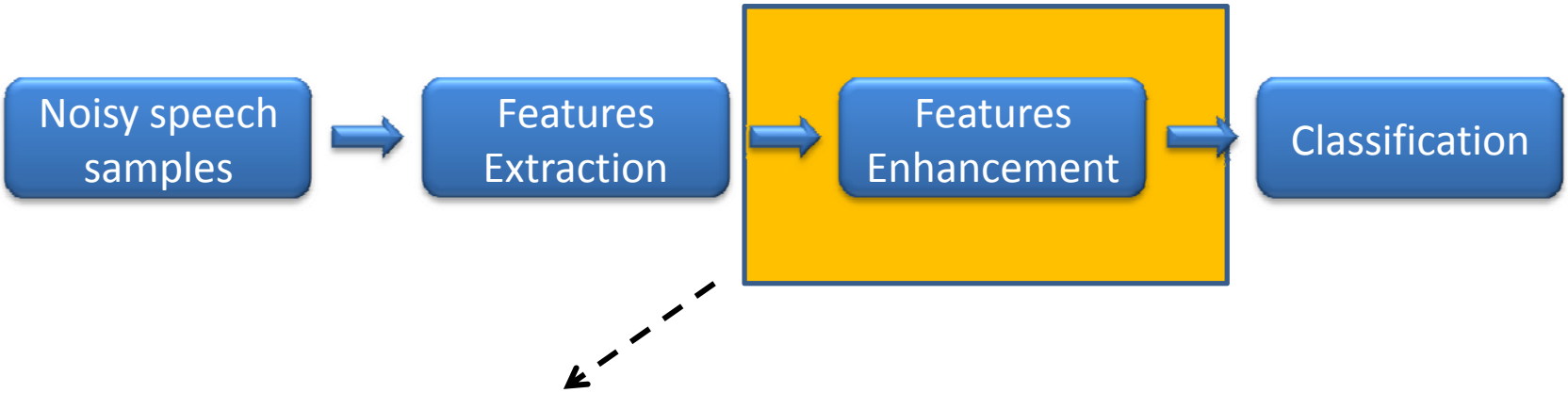


General noise filtering, on time domain signals[2]

[1] Nadav Merlis, Liora Neeman and Prof. Koby Crammer, "Hebrew Speech Recognition for iPhone", SIPL 2011

[2] I. Cohen and B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments", Signal Processing, Vol. 81, No. 11, Nov. 2001, pp. 2403-2418.

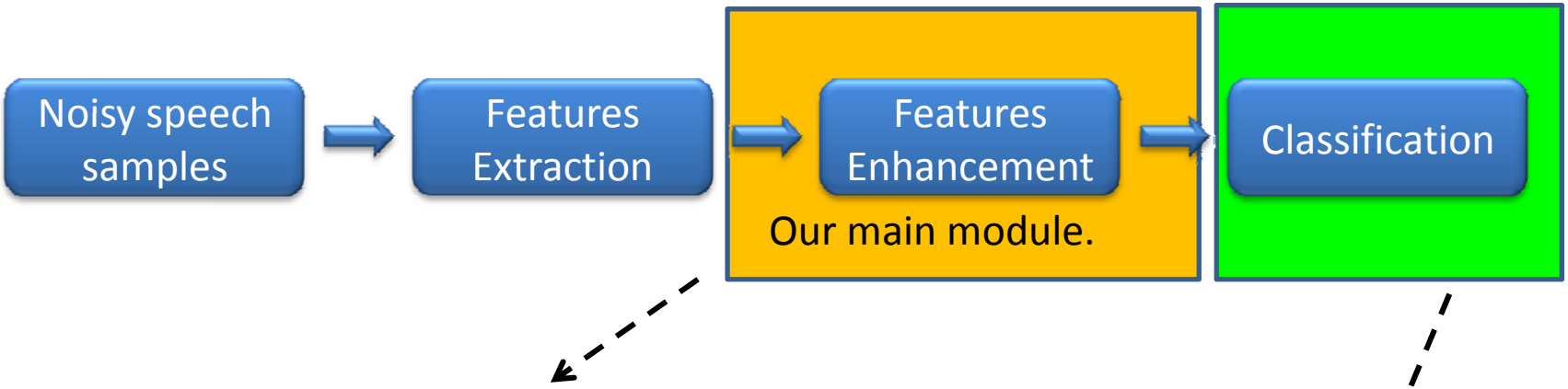
Proposed Solution #2



- Filtering in features domain[1]
- Based on statistical models for the speech and noise signals

[1] R. Haeb-Umbach and J Schmalenstroeer, “A comparison of particle filtering variants for speech feature enhancement”, Proc. of Interspeech, 2005

Our Proposed Solution

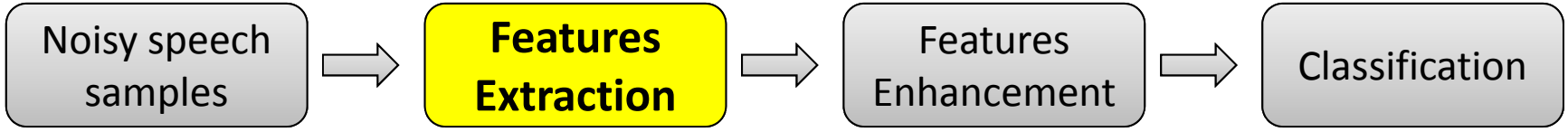


- Filtering in features domain[1]
  - o Bias consideration
  - o Smart re-sampling

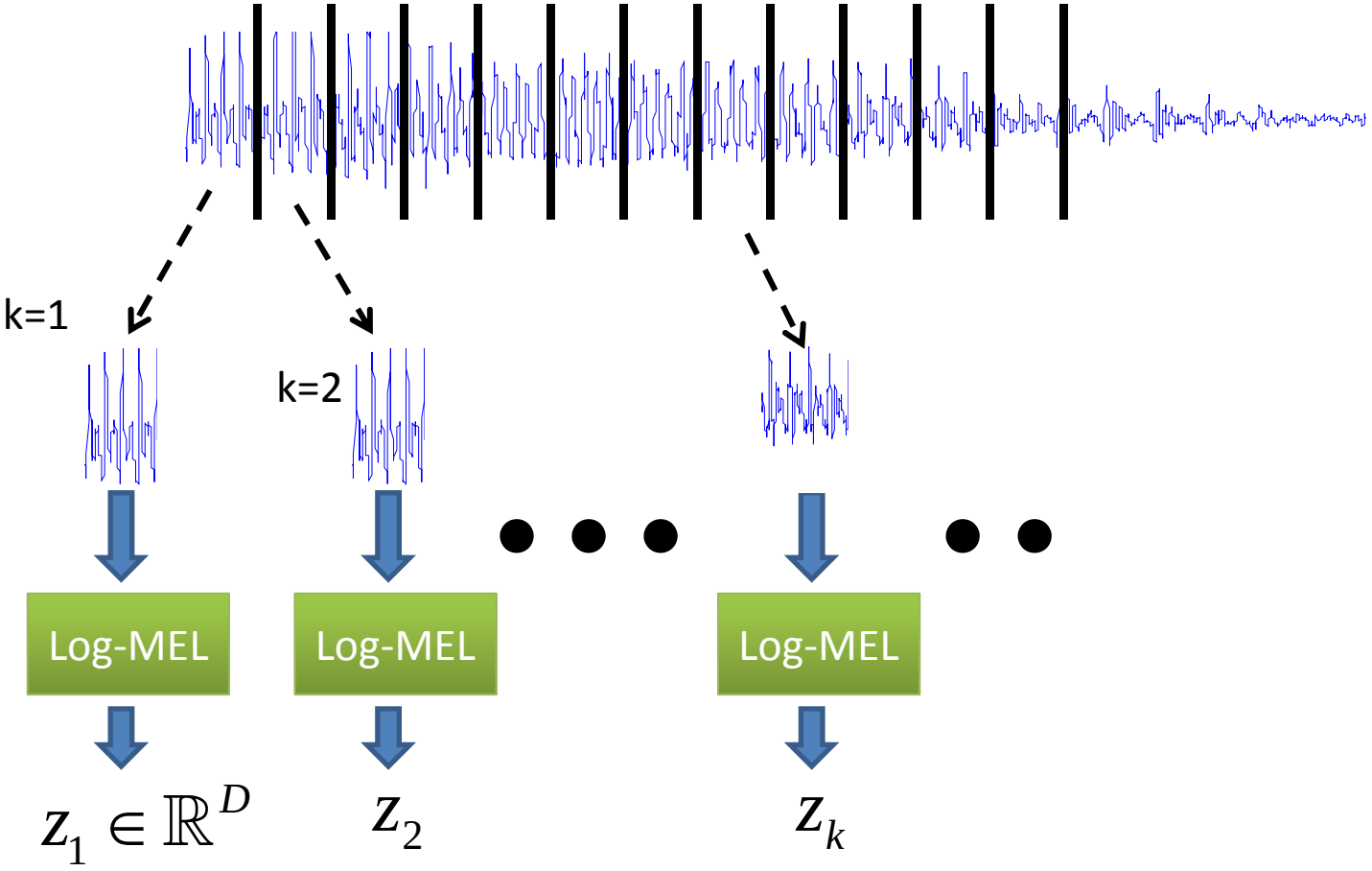
- o Adaption to our Features Enhancement system
- o Evaluation using max posterior

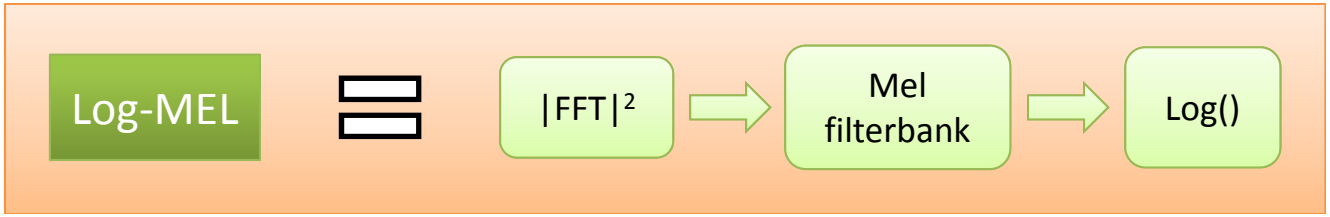
Speech Enhancement for Speech Recognition using Particle Filtering

The Features



Features Extraction





Features Extraction

Notations:

z_k - Noisy sample (at frame # k)

s_k - Clean speech

x_k - Noise

Resulted Equation:

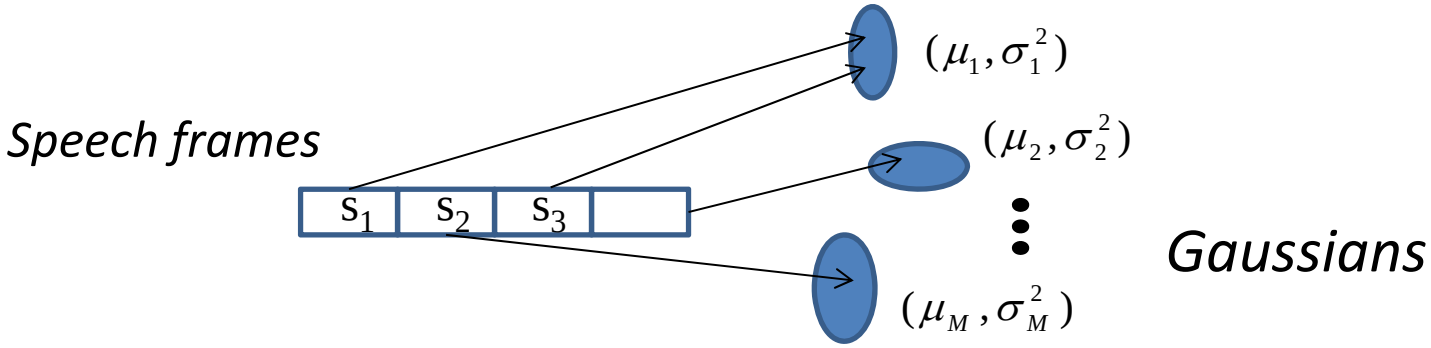
Assuming additive noise in time domain

$$z_k = s_k + \log(1 + e^{x_k - s_k})$$

Features model

speech features:

assumed to be drawn from a **Gaussian Mixture Model (GMM)**.



Noise model:

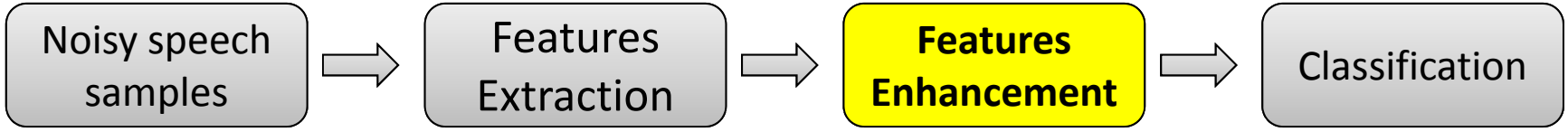
"environmental noises" ⇔ Correlation between frames exist

First order Auto Regressive (AR) Process

$$x_k = A \cdot x_{k-1} + w_k$$

Speech Enhancement for Speech Recognition using Particle Filtering

Enhancement Module



Estimation Problem

- **Input:**

  Non-linear State System:

$$x_k = A \cdot x_{k-1} + w_k$$

$$z_k = s_k + \log(1 + e^{x_k - s_k})$$

z_k - Noisy sample (at frame # k)

s_k - Clean speech

x_k - Noise

  Aim:

  Estimate (track) iteratively: $\hat{x}_k$ from samples- $z_{1:k} = (z_1, \ldots, z_k)$

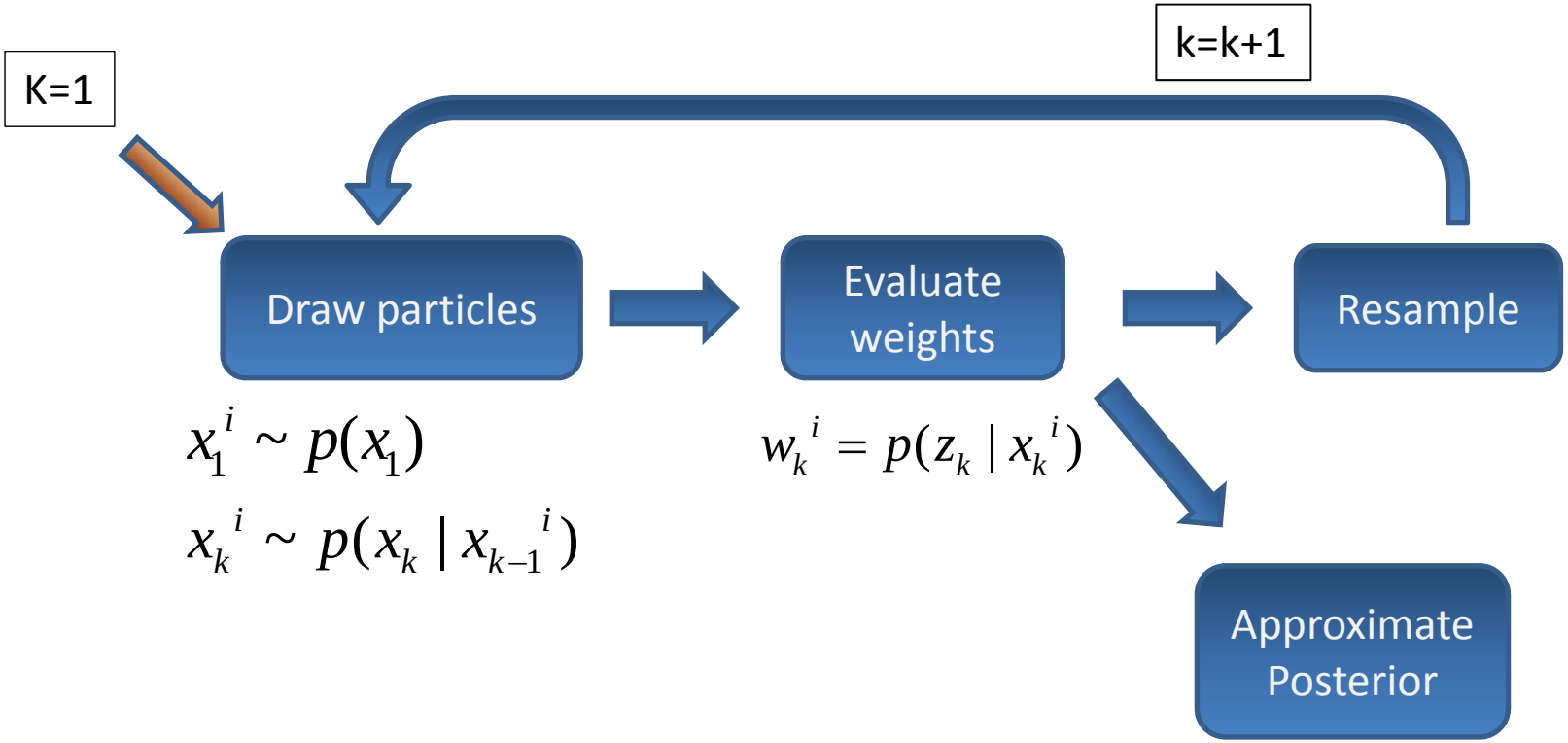
  Following, derive clean speech (s_k) estimation

The state system is highly non-linear => Kalman filter won't work

- ***Solution:*** **Particle Filter (PF)**

  Monte Carlo algorithm for sequential estimation

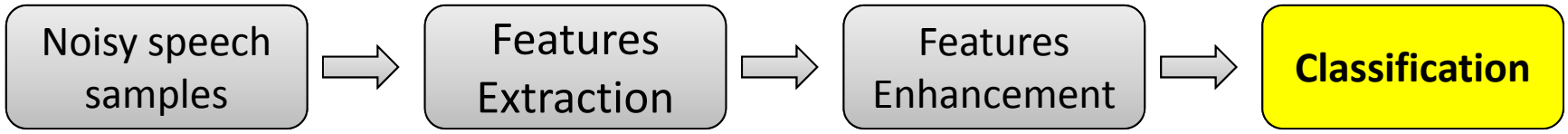
Particle Filter



$$\hat{p}(x_k \mid z_{1:k}) = \sum_i w_k^i \cdot \delta_{(x_k - x_k^i)}$$

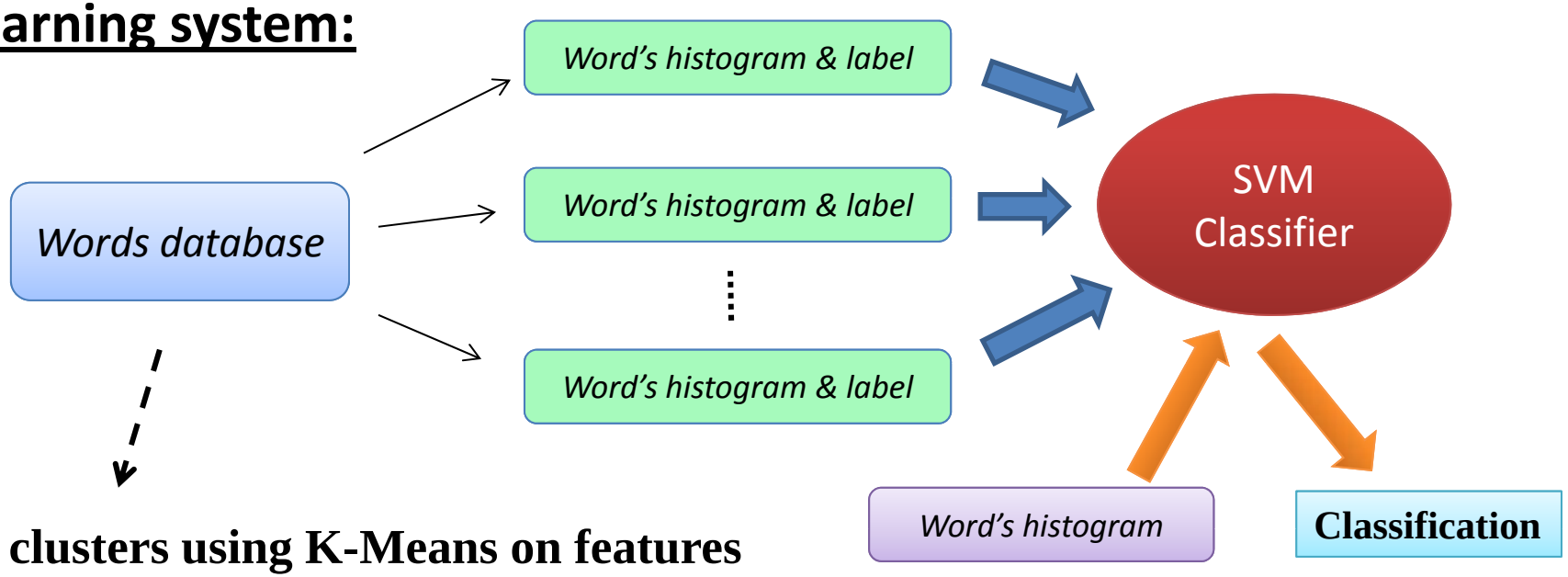
Speech Enhancement for Speech Recognition using Particle Filtering

Classification Module



Speech Recognition system*

A Learning system:



Train clusters using K-Means on features

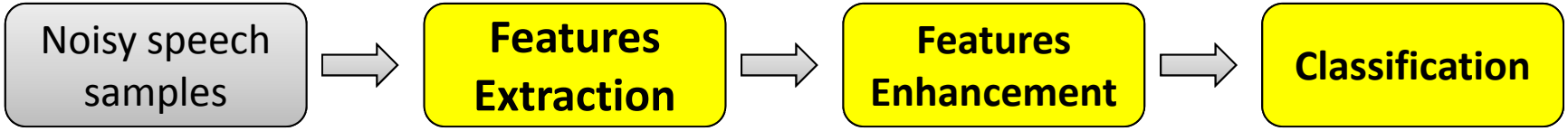
For each word:

- Associate each speech frame with cluster
- Create histogram for occurrences of clusters along each word

*Prof. Koby Crammer, Implemented by Nadav Merlis and Liora Neeman

Speech Enhancement for Speech Recognition using Particle Filtering

Our Main Improvements

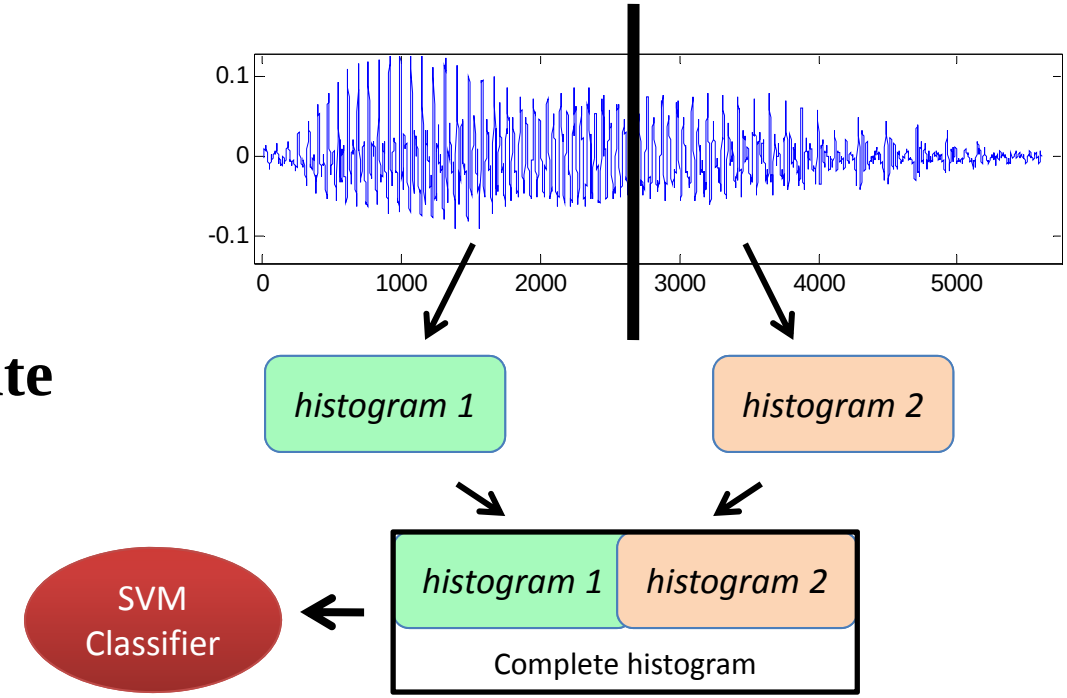


Improvement #1

Enhanced Speech Recognition system

- Using GMM *(instead of simple clustering)*
  - **Advantages:**
    - Introduces covariance
    - Adjusted to the speech model we use in the Particle Filter (see next...)

- Word division:

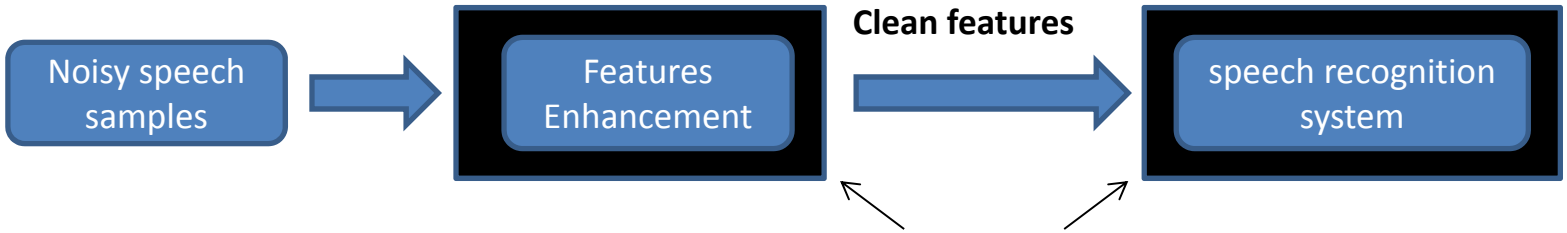


increases success rate by at least 5%

Improvement #2

Max Posterior Estimation

Direct approach:



Problem: Filter can't be ideal

Optimal solution:

Choose Gaussians by Max Posterior:

$$\hat{m}_k = \arg\max_{m_k} \{ p(m_k \mid z_{1:k}) \} = f\left(p(x_k \mid z_{1:k}) \right)$$

Gaussian Index at K'th frame

Evaluate using the particle filter results:

$$\hat{p}(x_k \mid z_{1:k}) = \sum_i w_k^{\,i} \cdot \delta_{(x_k - x_k^{\,i})}$$

Improvement #3

Bias Reduction

- AR model is adjusted to zero mean signals.
- The noise features are generally not zero mean. $E[X_k] = c \neq 0$

Our solution

$$z_k = s_k + \log(1 + e^{x_k - s_k})$$

$$z'_k \triangleq z_k - c \;,\; s'_k \triangleq s_k - c \;,\; x'_k \triangleq x_k - c \quad \Rightarrow \quad z'_k = s'_k + \log(1 + e^{x'_k - s'_k})$$

1) Estimate noise mean- c.

2) Decrease from samples- $z'_k \triangleq z_k - c$

3) Decrease from the speech Gaussians means- $\mu'_m \triangleq \mu_m - c$

4) Increase estimation- $\hat{s}_k \triangleq \hat{s}'_k + c$

Improvement #4

Improved Sampling

- Recall that: $z_k = s_k + \log(1 + e^{x_k - s_k})$
  - The noise must be smaller than the noisy speech
- Some of the particles might have zero weights:

$$w^i_k = p(z_k \mid x^i_k) \Big|_{x^i_k \geq z_k} = 0$$

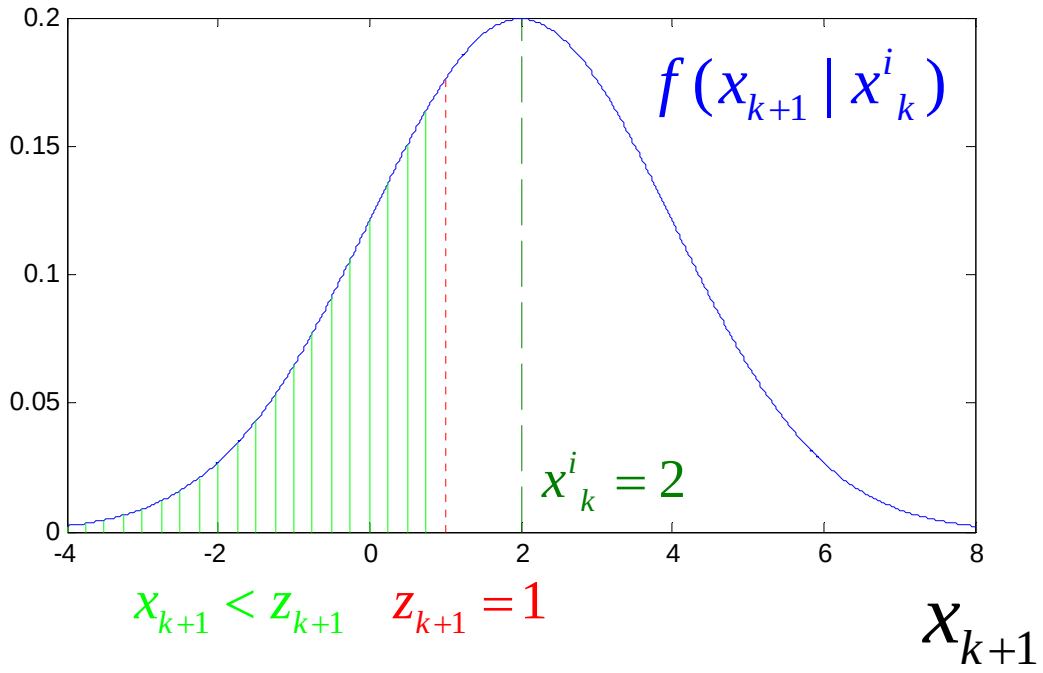
- A zero weight particle is not effective
- Reduced number of effective particles => worse estimation!
- Sometimes **ALL** particles receive zero weight…

Improvement #4

Improved Sampling

Our solution

sample in available region



- Draw only from green part
- Set initial weight: $w^i_{(initial)} = p(x^i_{k+1} < z_{k+1} \mid x^i_k)$

Speech Enhancement for Speech Recognition using Particle Filtering

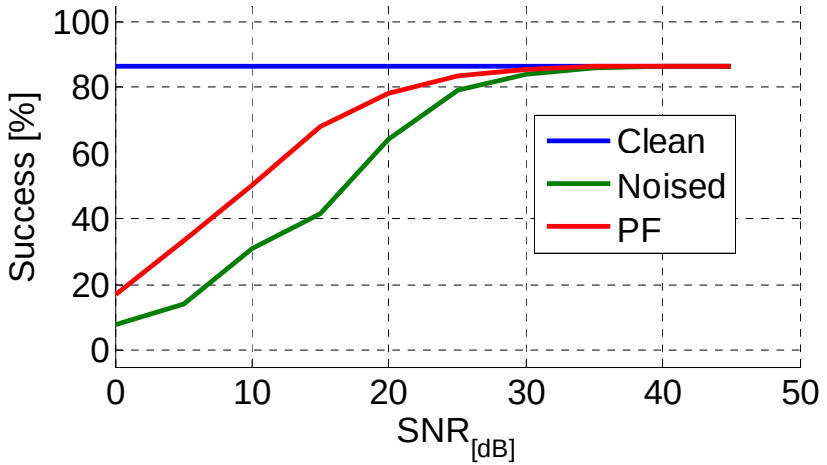
Results

Results – Preface

- The results are based on cross-validation over the entire database (ISOLET).
- Results show success rate per SNR.

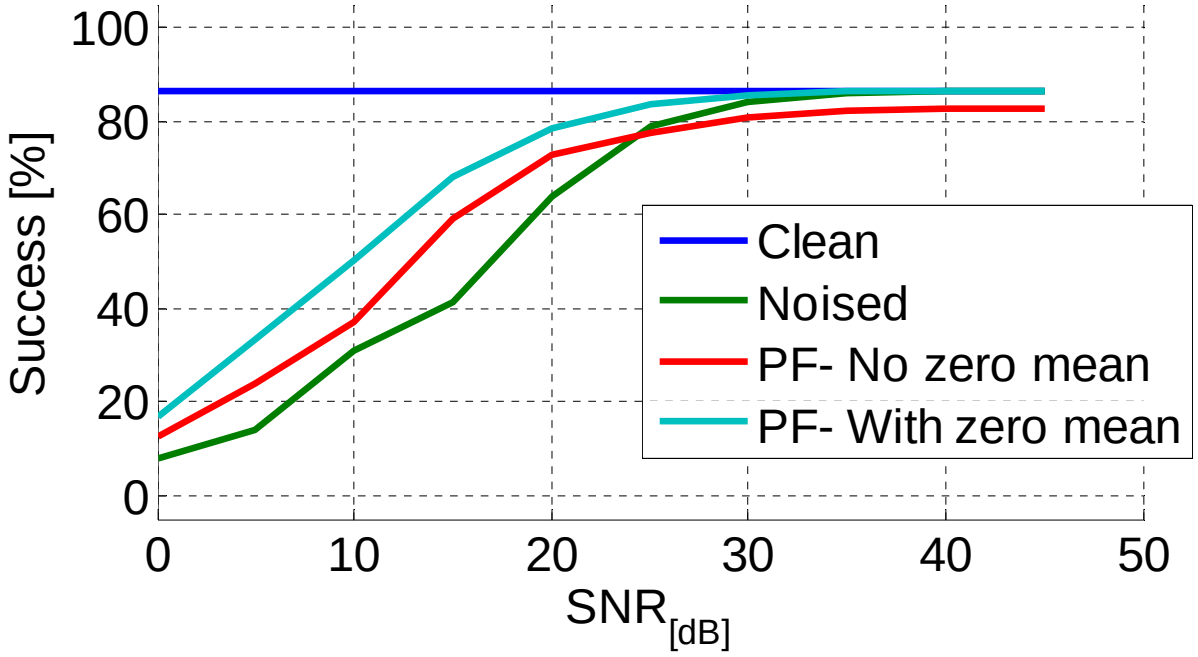
- ‘Clean’ – achieved success rate without noise.
- ‘Noised’ – achieved success rate without using any filter.

Sample results:



Implementation difference

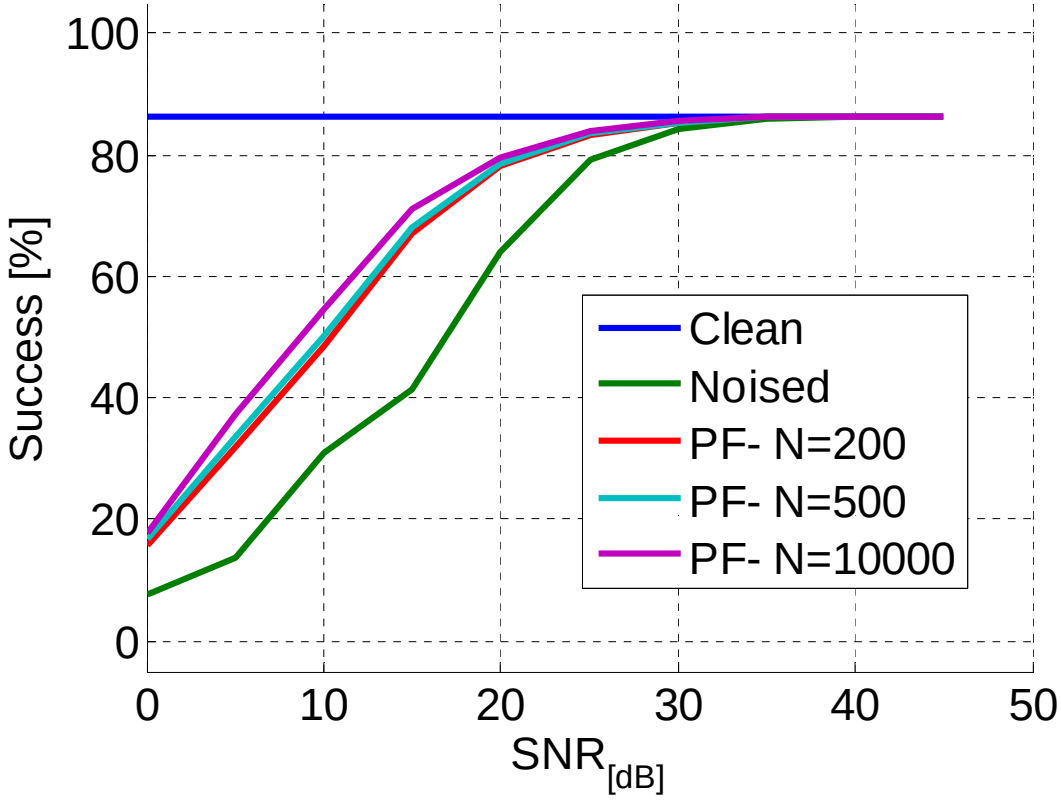
AR Adjustment:



- Significant improvement is achieved when decreasing the noise estimated mean

Particle Filter- Parameters

Particles Number:



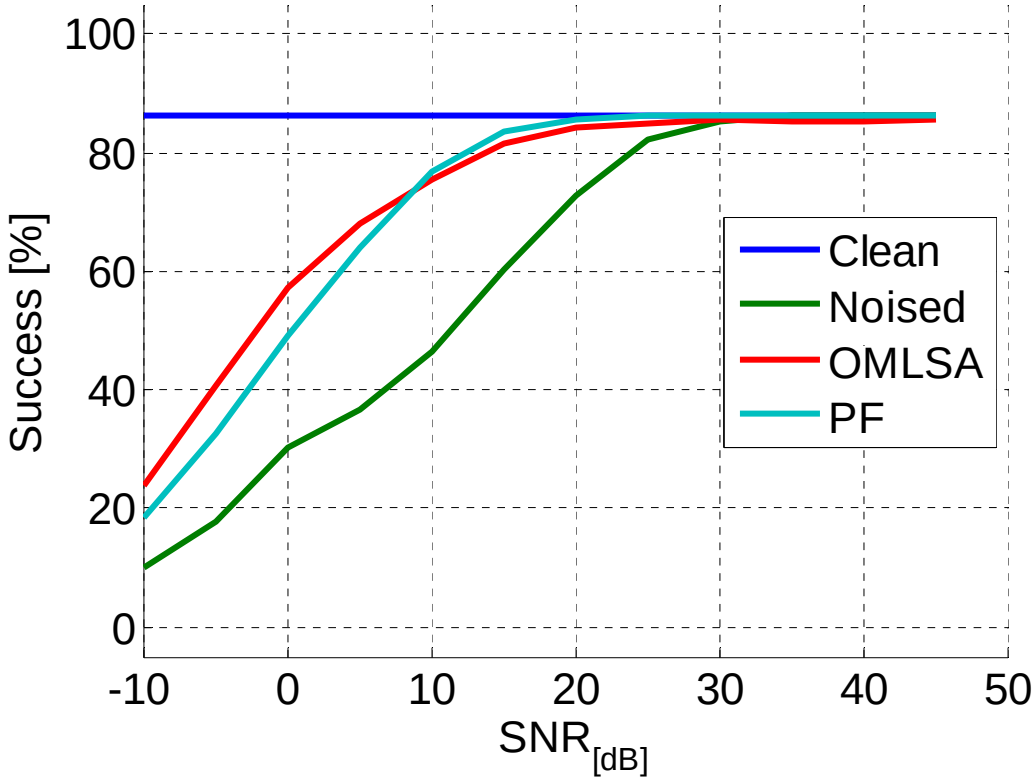
- Obvious improvement as the particles number increase.
- Note: Computation time is linear in the particles number.

Comparison

- Comparison to alternative- using OMLSA Filter on time domain samples

Tank Noise:

Stationary and slow changing

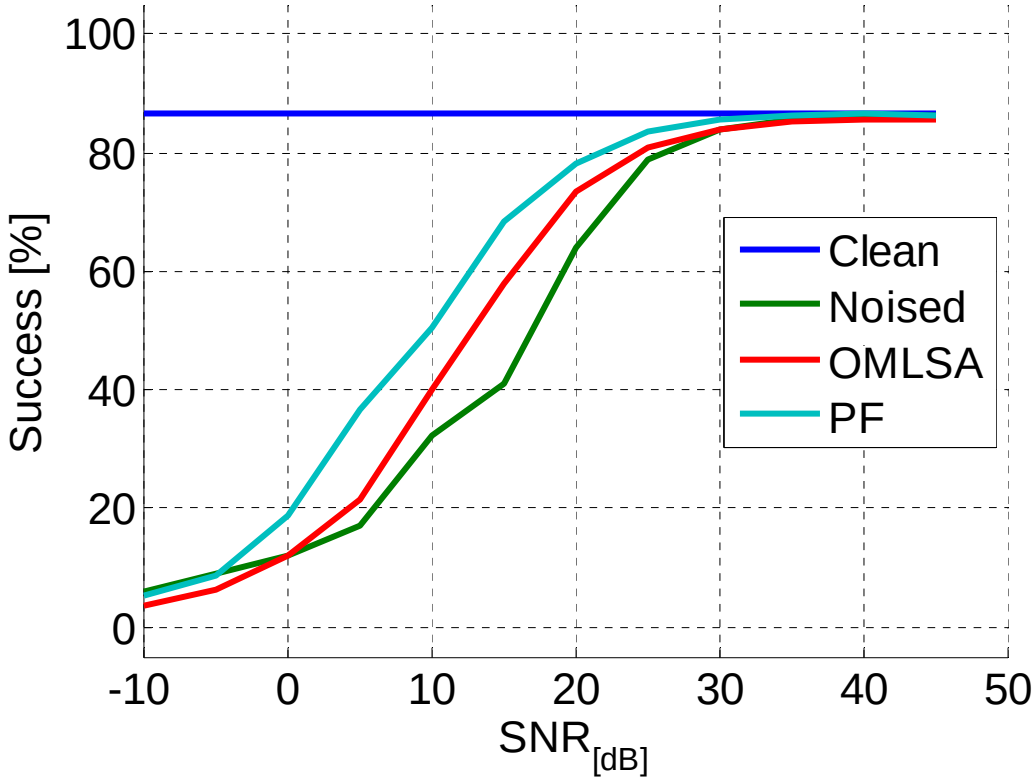


Comparison

- Comparison to alternative- using OMLSA Filter on time domain samples

Babble Talk Noise:

Stationary and rapidly changing signal

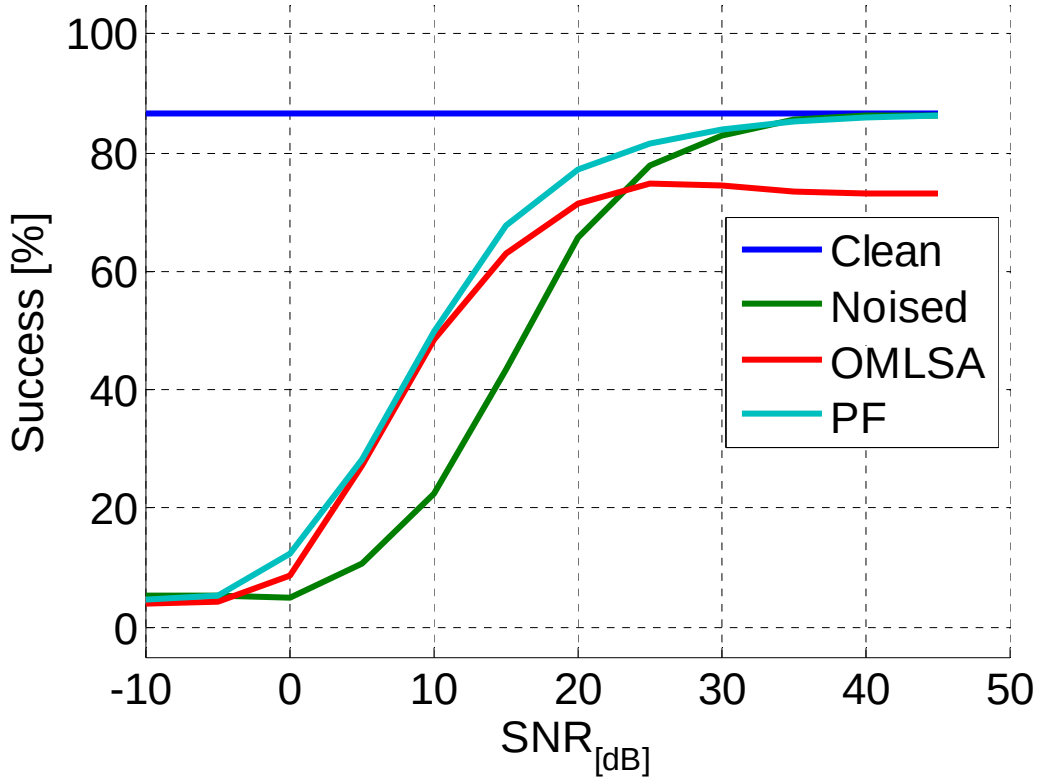


Comparison

- Comparison to alternative- using OMLSA Filter on time domain samples

Laugh Noise:

Not stationary



Speech Enhancement for Speech Recognition using Particle Filtering

Summary

Summary

- We used two Building Blocks:
  - Speech Recognition system
  - Enhancement in features domain.
- Introduced our improvements:
  - Split histograms - Max posterior estimation
  - Bias reduction - Improved sampling
- **The Results:**
  - Great improvement (up to 30%) compared to non-filtered signals
  - Significant improvement (up to 20%) over using the OMLSA filter, especially when the noise doesn't fit its assumptions

What Could Be Done Next?

- Models improvement:
  - Introduce correlation between speech frames
  - Time Varying AR
    - Continually varying of parameters
    - Different sets of parameters (mainly different bias).
- Improve the speech recognition:
  - Use the inter-frame dependency

Speech Enhancement for Speech Recognition using Particle Filtering

The End