



Missing Data Solutions (MIDAS)

Hugo Van hamme¹, Yujun Wang¹, Jort Gemmeke², Bert Cranen², Rudi Vuerinckx³

(¹) K.U.Leuven –ESAT (²) Radboud Universiteit Nijmegen – CSLT (³) Nuance Communications Belgium

<http://www.esat.kuleuven.be/psi/spraak/projects/MIDAS/>



1. Project Background

- MIDAS will extend “SPRAAK” automatic speech recognizer with robustness to non-stationary noise
- The *Missing Data Theory (MDT)*
 - if speech is masked by noise, it is considered as lost
 - humans are good at handling missing information, HMMs are not
 - MDT can handle non-stationary noise
- MDT-based recognition
 - reliable and robust spectral mask estimation, need a *Missing Data Detector (MDD)*
 - discard the masked (green) information
 - reconstruct it from the reliable (blue) information + speech model (imputation) for all recognition hypotheses
 - allow any value with probability given by the speech model (marginalization)

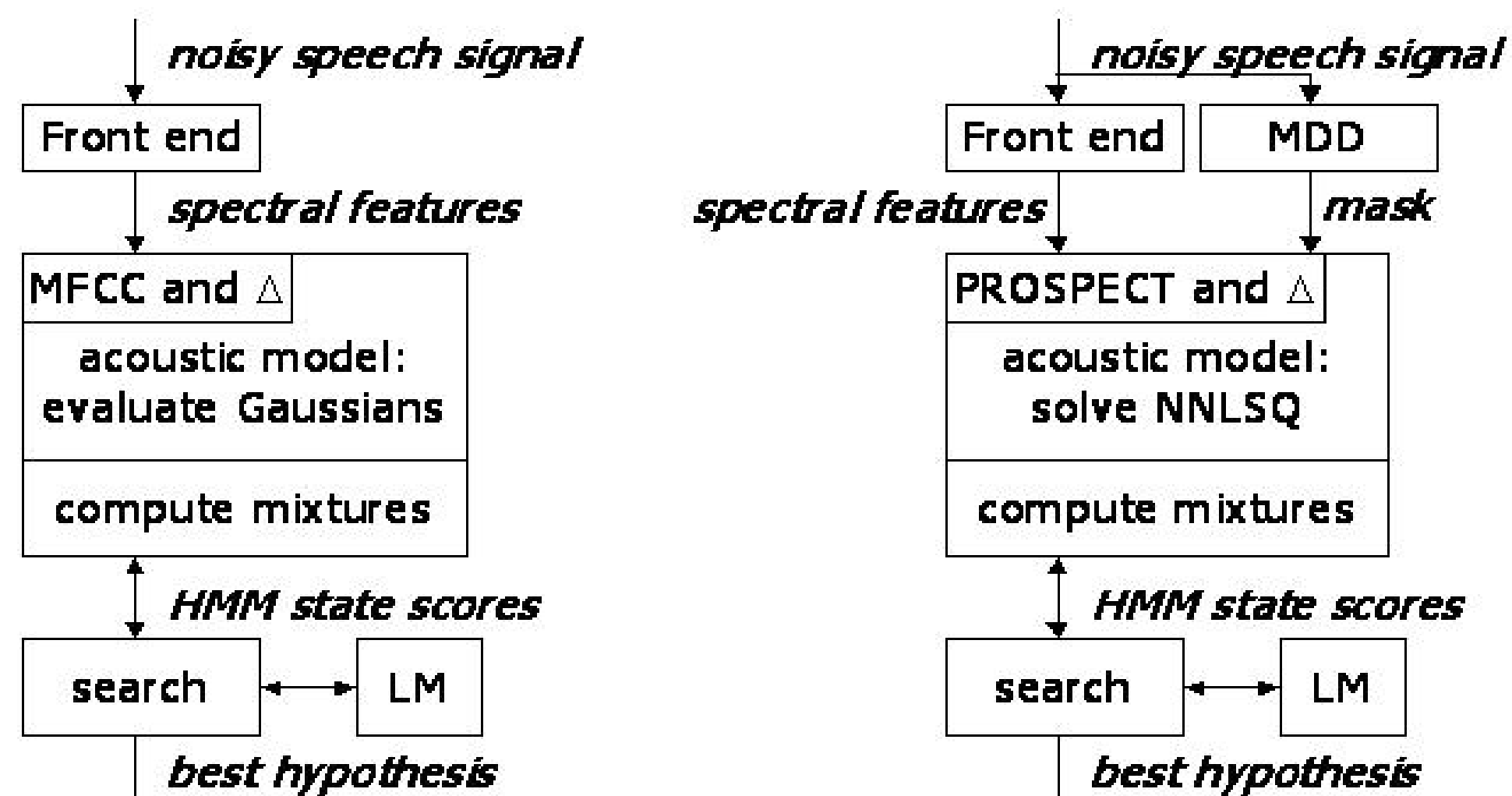
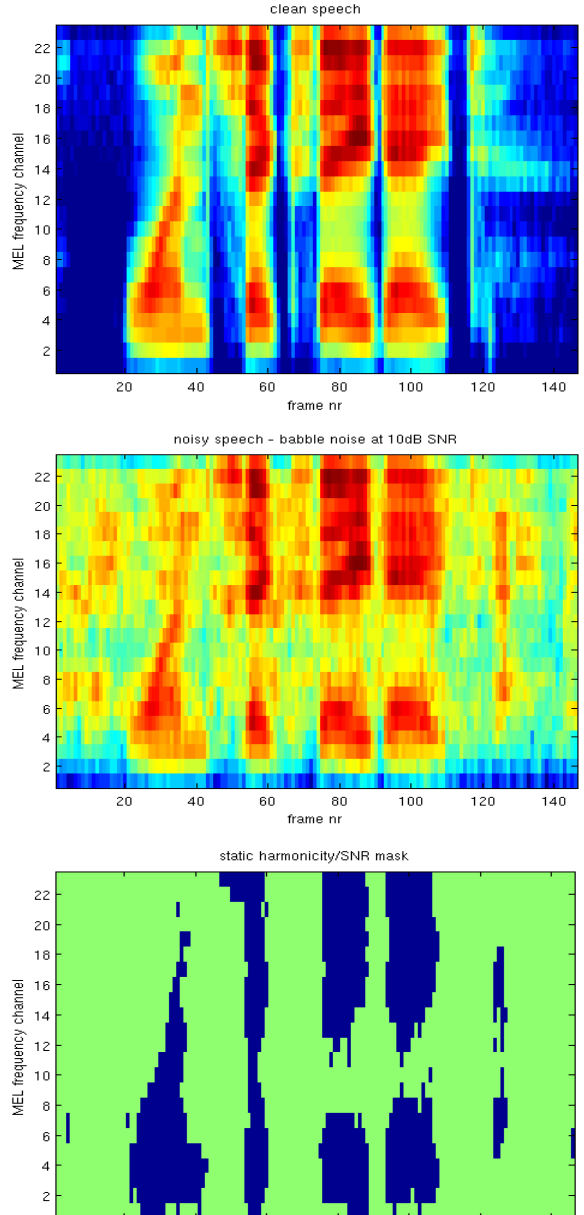


Figure 1: a typical HMM-based ASR system using cepstral and Gaussian mixtures. LM = language model. Δ = computation of time derivatives.

Figure 2: missing data system. The Missing Data Detector (MDD) is added and evaluation of Gaussians is replaced by solving a NNLSQ problem.

Clean speech, noisy speech and mask **Architecture of a “traditional” HMM recognizer (left) and a MDT-based recognizer (right)**

2. Project Deliverables

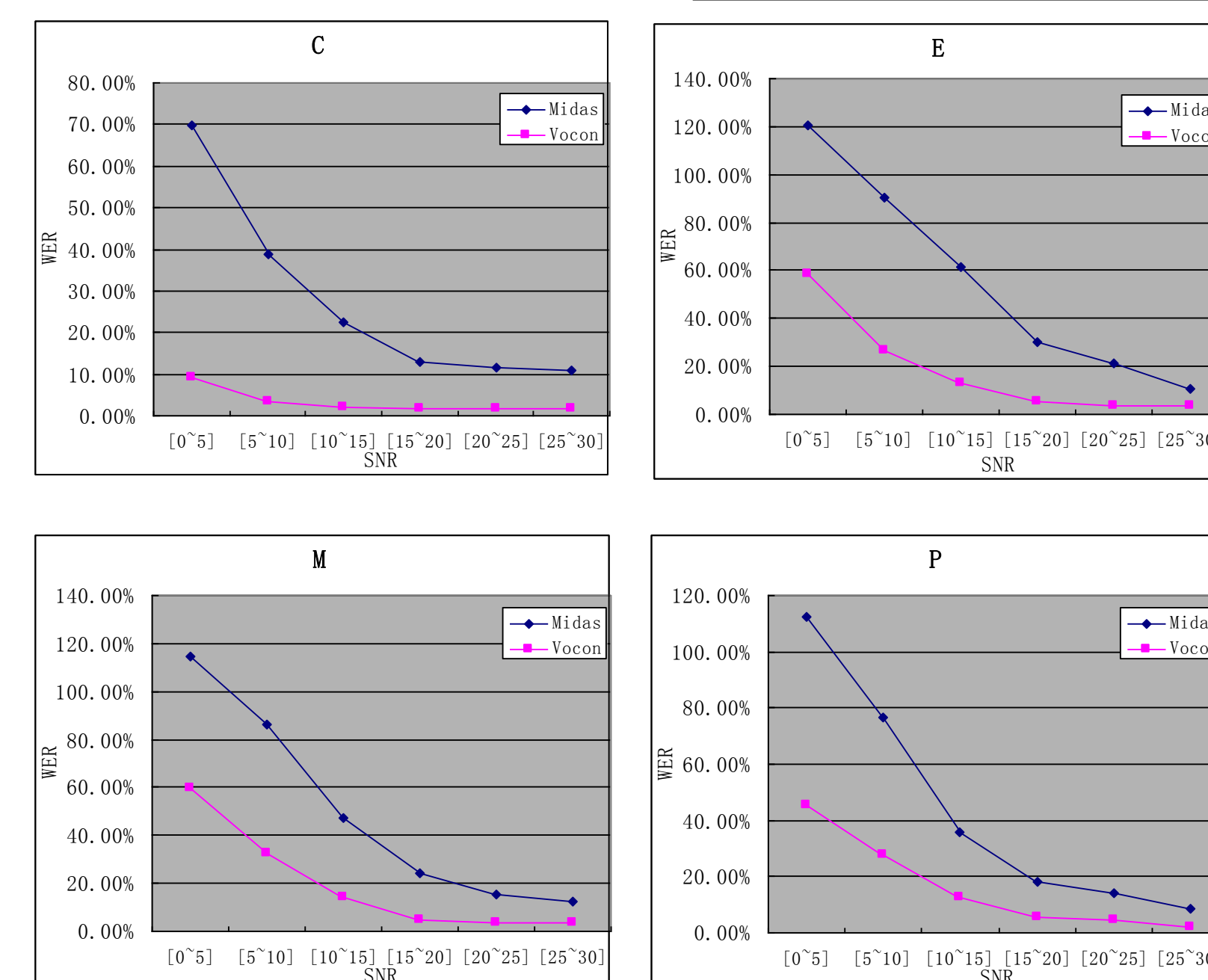
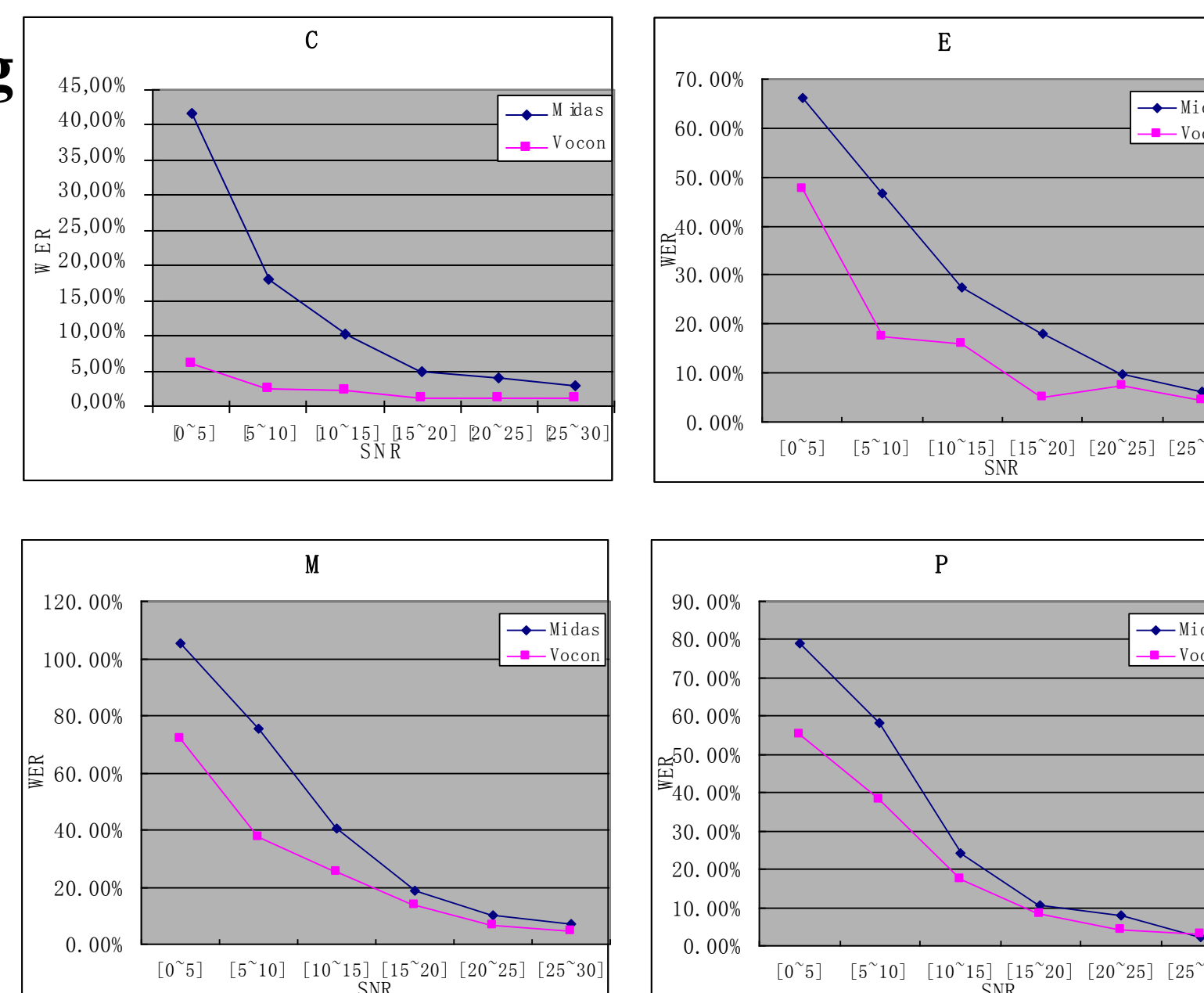
- D1: Training and test material defined (done).
- D2: Benchmark report of state-of-the-art system (done).
- D3: Benchmark report of baseline MDT system
 - “connected digits” and “isolated words” are tested whereas “when”
 - “spelling letter” and “natural number” are to be tested
- D4: MDT-based recognizer version 2
 - Improving binary masks (done)
 - Imputation using wider time-context (good progress)
 - Faster algorithms for NNLSQ (done by Gaussian selection, rapid solution required)
- D5: MDT-based recognizer version 3
 - Soft masks (done), soft masks used in back-end (done)
 - Sound class dependent masks (good progress)
- D6: Final MDT system integrated in SPRAAK
- D7: Benchmark report of final version

3. MIDAS and VOCON 3200 benchmark

Connected digits testing results: MIDAS vs. Vocon 3200 over each noise type

Noise types:

- C: Car
- E: Entertainment
- M: Office
- P: Public Hall



Isolated word testing results: MIDAS vs. Vocon 3200 over each noise type

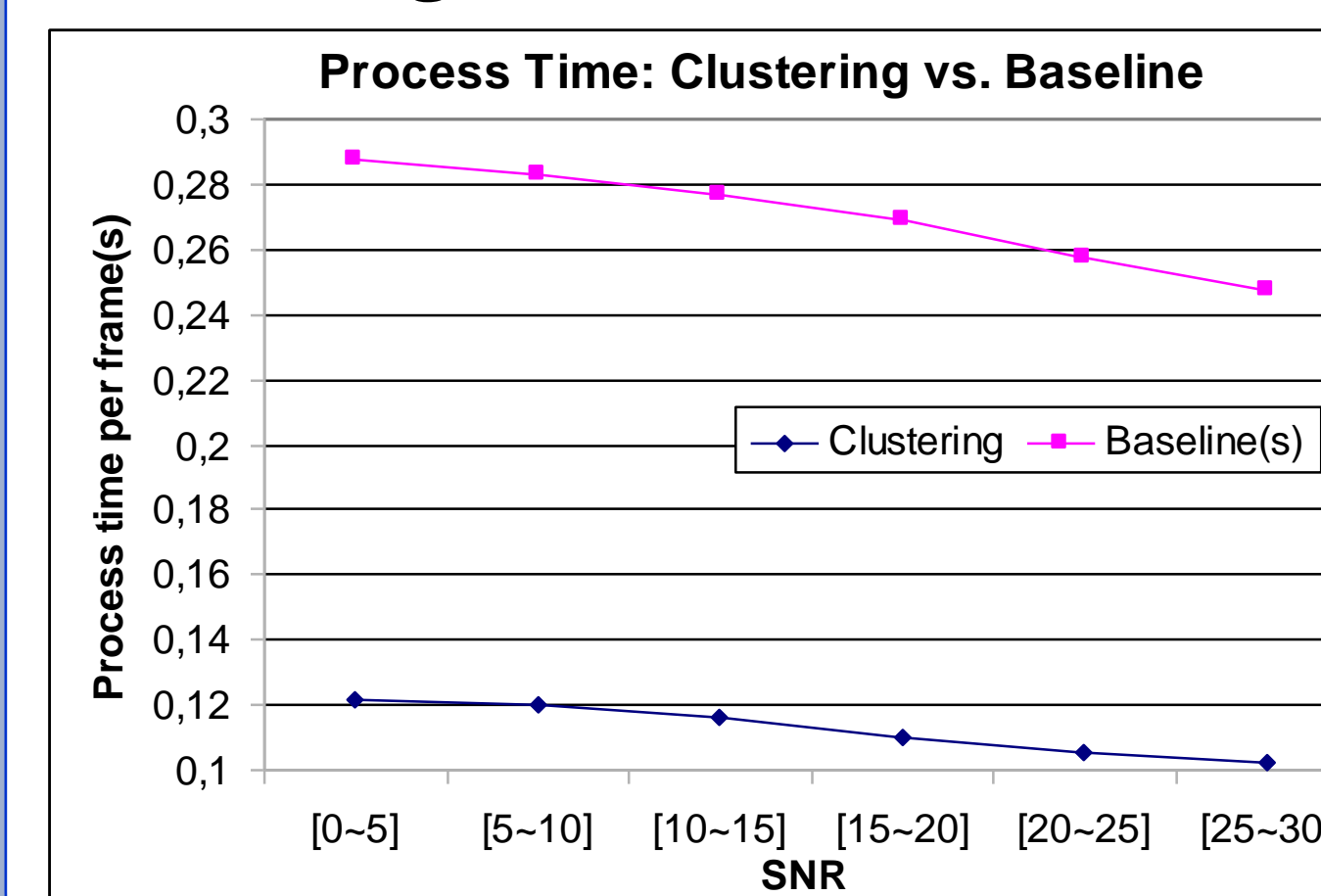
Conclusions: There is still big performance gap between MIDAS recognizer and Vocon 3200

4. Speed-up by Gaussian selection

Problem: CPU-time dominated by Gaussian evaluation with MDT => Avoid computation of Gaussians

- **baseline:** compute all Gaussians
- **cluster:** cluster Gaussians with KL-criterion; compute all Gaussians of N-best scoring clusters

The testing results on MIDAS Isolated words task



Conclusion: cluster method is more efficient without recognition accuracy lost

4. Improved Binary/Fuzzy Masks

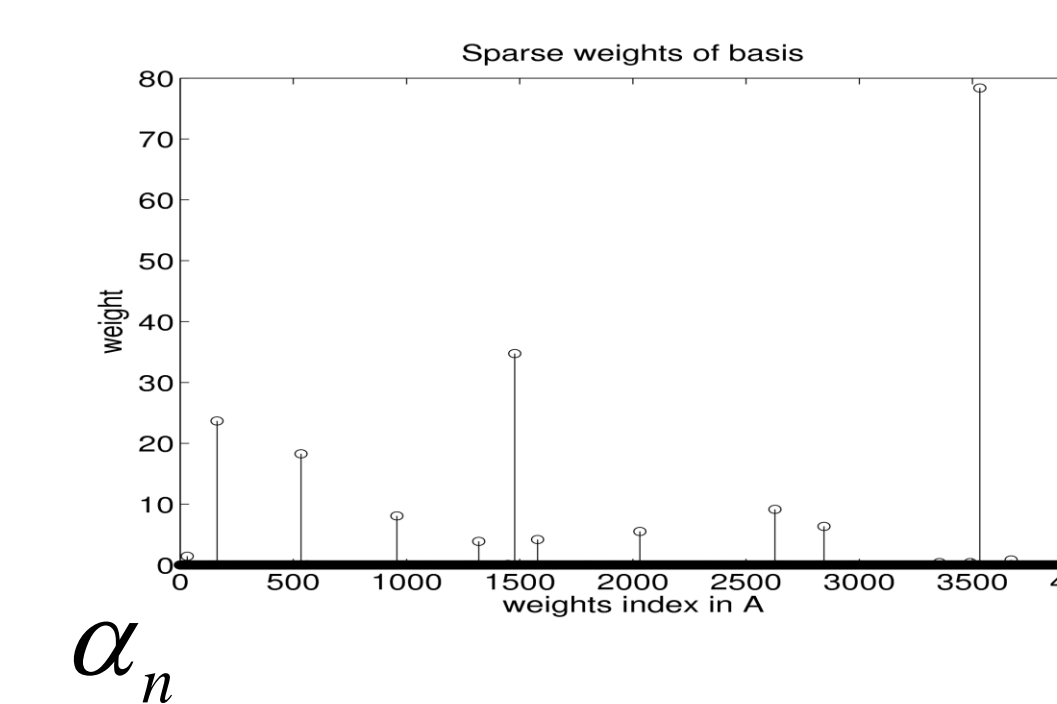
- Use Support Vector Machines (SVMs)
- Learn to distinguish reliable and unreliable areas using Noisy speech
- Combine multiple features:
 - Harmonic decomposition
 - Local shape of the spectrum (flatness)
 - Noise energy floor per frequency band
 - Noisy speech features
- Optionally provides probability measures (*fuzzy masks*)
- Extension to *state-dependent* masking (one SVM for every HMM state)
- Good results on AURORA-2, now investigating MIDAS data

4. Imputation using wider time-context

- MDT ASR performance decreases at low SNRs (below 0 dB)
- Possibly due to frame-by-frame processing: Many frames do not contain *any* reliable features
- Solution: Use wider time-context
- Approach: Find a *sparse* representation in a set of example speech signals using only the reliable areas of the spectrum.

$$\vec{y} = \sum_{n=1}^N x_n \vec{a}_n = \mathbf{A} \vec{x}$$

$$\min \|\vec{x}\|_1 \text{ subject to } \vec{y}_r = \mathbf{A}_r \vec{x}$$



$$\vec{\tilde{y}} = \begin{bmatrix} \vec{y}_r \\ \vec{y}_u \end{bmatrix} = \begin{bmatrix} \vec{y}_r \\ \mathbf{A}_u \vec{x} \end{bmatrix}$$

