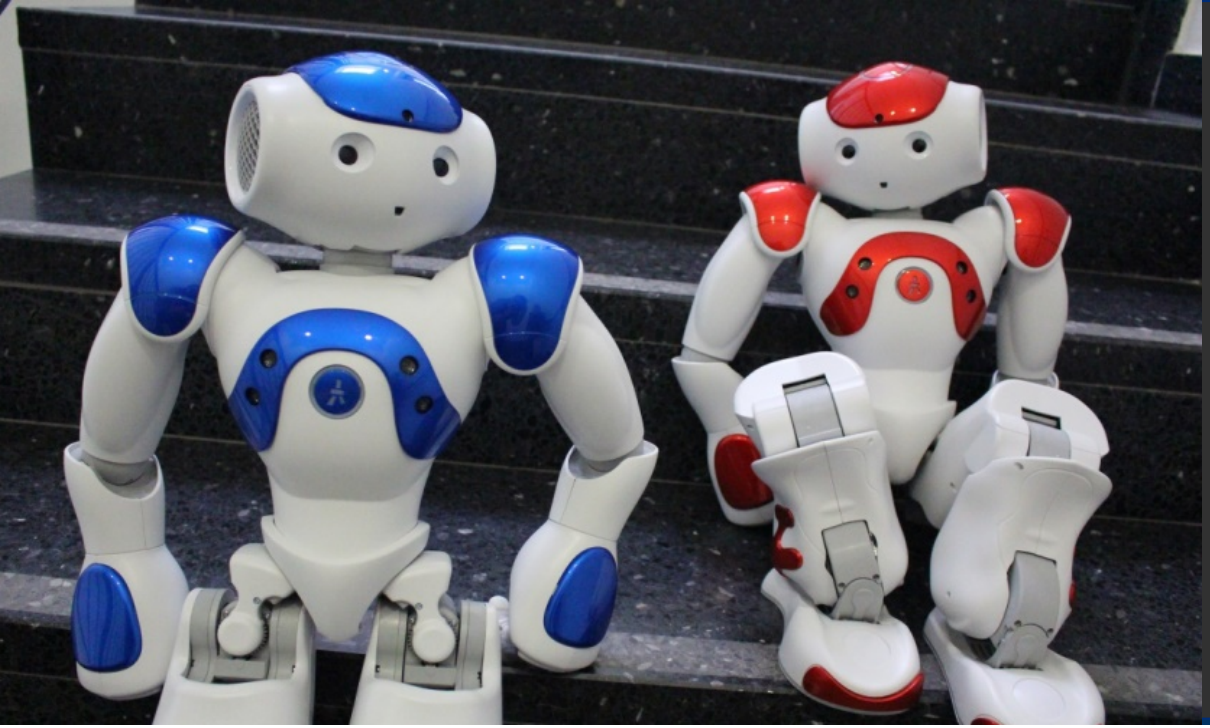


TWO!EARS

Perception in the real world: Bayesian and active approaches to environmental robustness

Cognitive Signal Processing Group
Dorothea Kolossa

Fakultät Elektrotechnik und Informationstechnik, Ruhr-Universität Bochum,
“Physics of Hearing” Workshop @ Kavli Institute of Theoretical Physics, UCSB

RUB

Cognitive Signal Processing Group

Research Associate:
Steffen Zeiler

PhD students:
Benedikt Bönninghoff
Hendrik Meutzner
Christopher Schymura
Mahdie Karbasi
Dennis Orth
Lea Schönherr
Former Scientists
Ahmed Hussen Abdelaziz
Sebastian Gergen

Student Research Assistants:
Juan Diego Rios Grajales
Jan Hünnemeyer
Tobias Isenberg
Diana Castano Marin

with many thanks to our collaboration partners...

- Bucknell University, Lewisburg, PA (Prof. Robert Nickel)
- LAAS-CNRS, Toulouse, France (Prof. Patrick Danes)
- Mitsubishi Electric Research Labs, Cambridge, MA (Dr. John Hershey)
- ICSI, UC Berkeley (Prof. Nelson Morgan, Dr. Ahmed Hussen Abdelaziz)
- INESC-ID, Lissabon, Portugal (Dr. Ramón Astudillo)
- Aalto University, Finnland (Dr. Rahim Saeidi)
- TU Berlin / TU Ilmenau, Germany (Prof. Alexander Raake)
- DTU Copenhagen, Denmark (Prof. Torsten Dau, Dr. Tobias May)
- Univ. Rostock (Prof. Sascha Spors, Fiete Winter)

Introduction & Overview

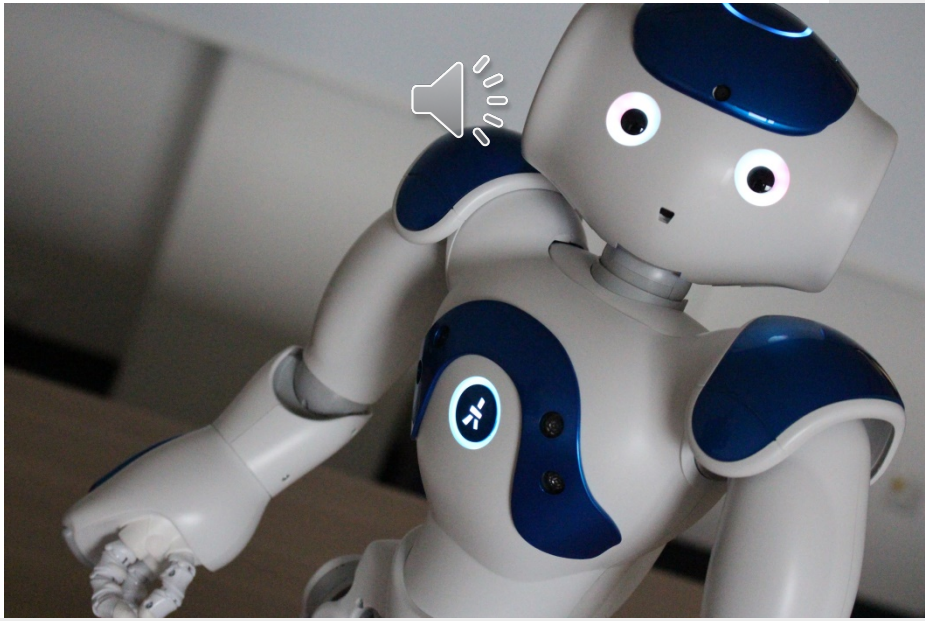
Machine listening

- environmental scene analysis
- source localization
- speech recognition

has made great progress in recent years.

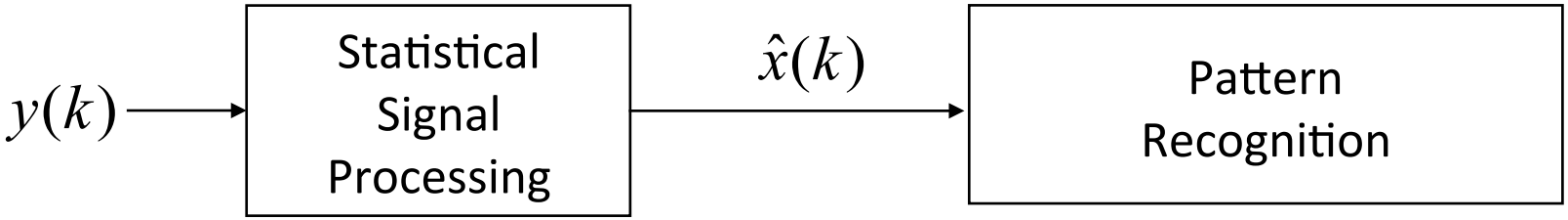
But reliability could still be improved in

- noisy environments
- reverberant environments
- and during periods of ego noise



Introduction & Overview

Signal processing approaches like noise reduction



[Boll1979] S. Boll: “Suppression of speech in noise using spectral subtraction” Proc. IEEE AASP-27, April 1979, pp. 113-120.

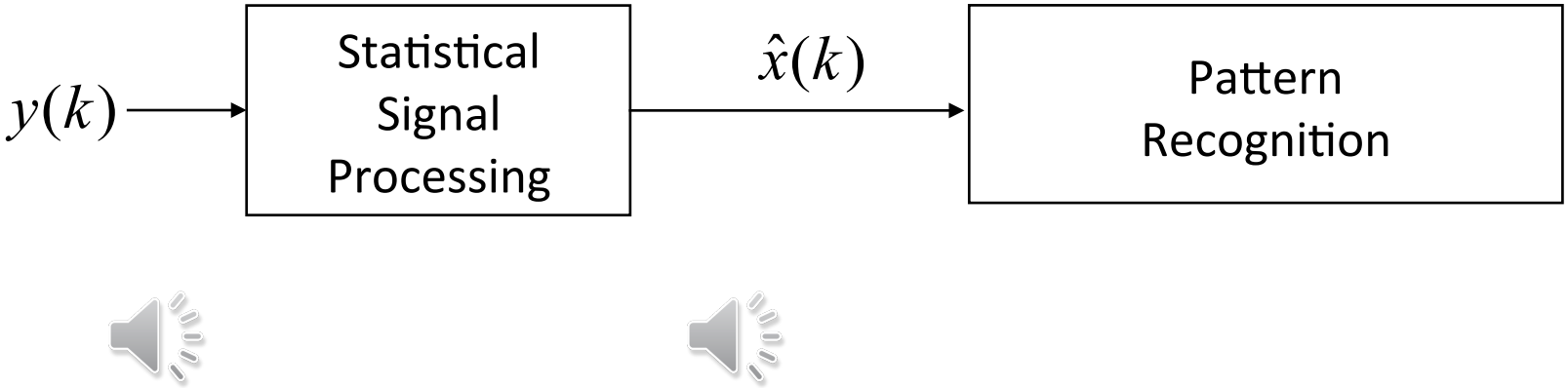
[Martin2003] Rainer Martin: “Statistical methods for the enhancement of noisy speech,” Proc. IWAENC 2003.

[LeRoux2012] Jonathan Le Roux, John R. Hershey: “Indirect model-based speech enhancement,” Proc. ICASSP 2012, pp. 4045–4048.

[LeRoux2013] Jonathan Le Roux, Shinji Watanabe, John R. Hershey: “Ensemble learning for speech enhancement”, Proc. WASPAA 2013.

Introduction & Overview

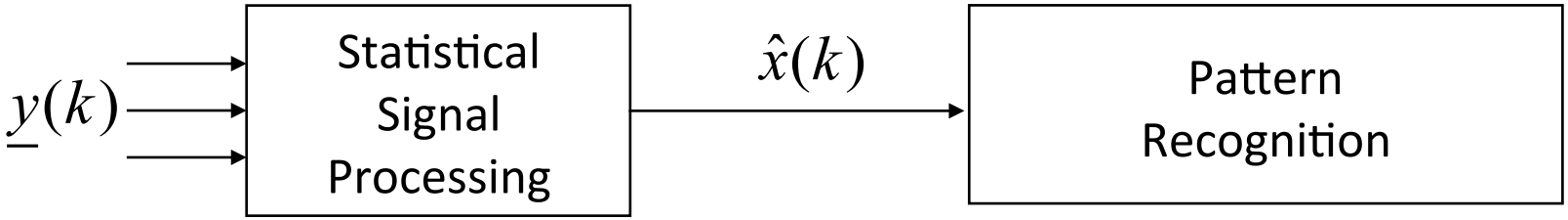
Signal processing approaches like noise reduction



[Taghia2016] J. Taghia, D. Kolossa, R. Martin: „ALE for Robots! A single-channel approach to robot self noise reduction“, IWAENC 2016.

Introduction & Overview

... or multichannel speech enhancement



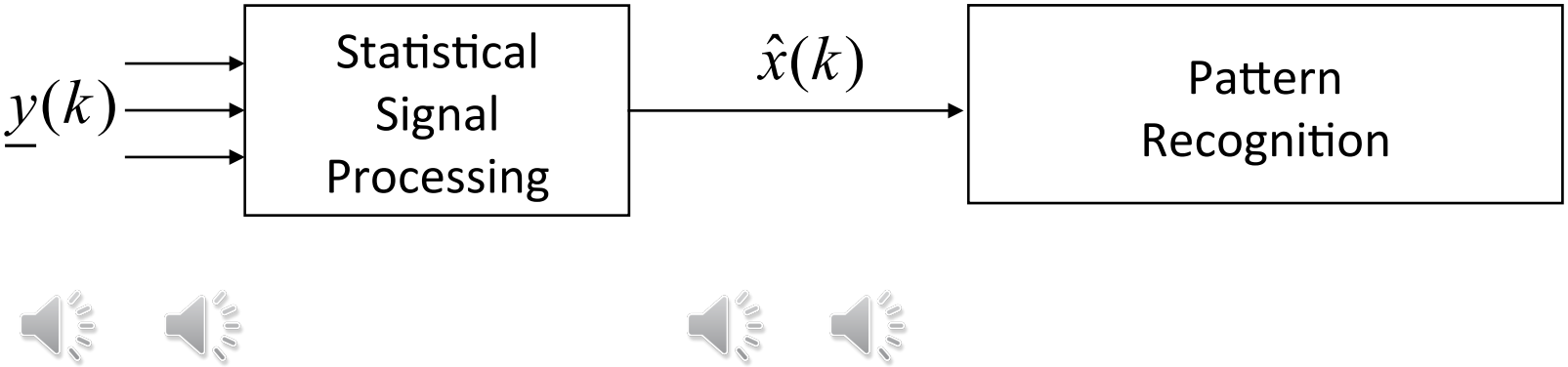
[Liutkus2011] A. Liutkus, R. Badeau, G. Richard: “Gaussian processes for underdetermined source separation” IEEE Transactions on Signal Processing, 59 (7), 3155-3167.

[Drude2015] L. Drude, A. Chinaev, R. Haeb-Umbach: “BLSTM-Supported GEV Beamformer Front-End for the 3rd CHiME Challenge,” Proc. ASRU 2015.

[Vincent] E. Vincent, T. Virtanen, S. Gannot: “Audio Source Separation and Speech Enhancement” to appear, Wiley.

Introduction & Overview

... or multichannel speech enhancement



[Kolossa2010] D. Kolossa, R. Fernandez Astudillo, E. Hoffmann and R. Orglmeister: „ Independent Component Analysis and Time-Frequency Masking for Speech Recognition in Multitalker Conditions“, EURASIP Journal on Audio, Speech, and Music Processing. vol. 2010, Article ID 651420, 13 pages, 2010.

help, but have a narrow interface to robot audition and typically provide no reliability information.

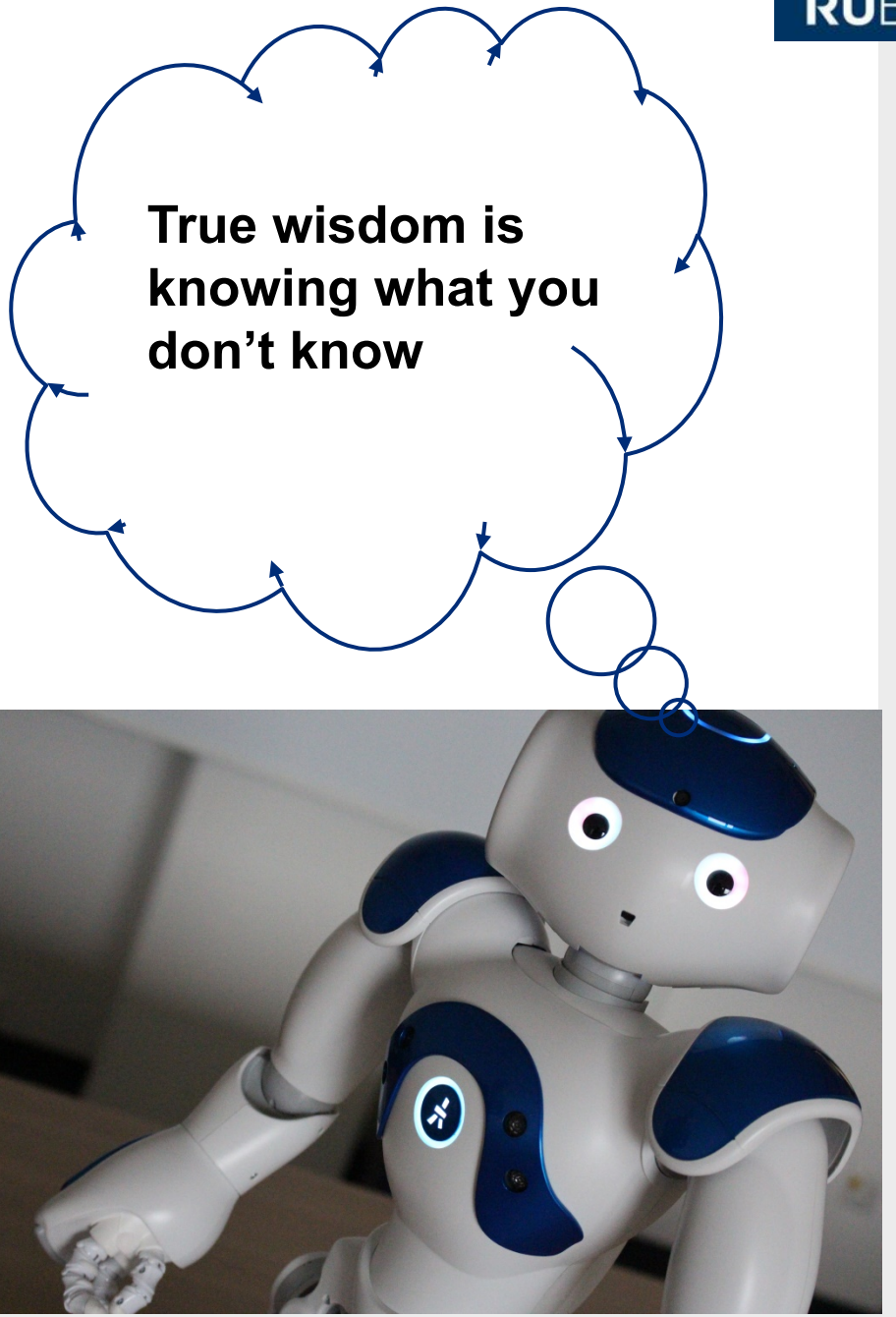
Introduction & Overview

Machine listening

Two major issues

1) Very fast adaptation

2) Tracking your own reliability



Introduction & Overview

Idea

Profit from human perception strategies to improve both aspects of machine listening (rapid adaptation, knowing what you don't know).

Outline

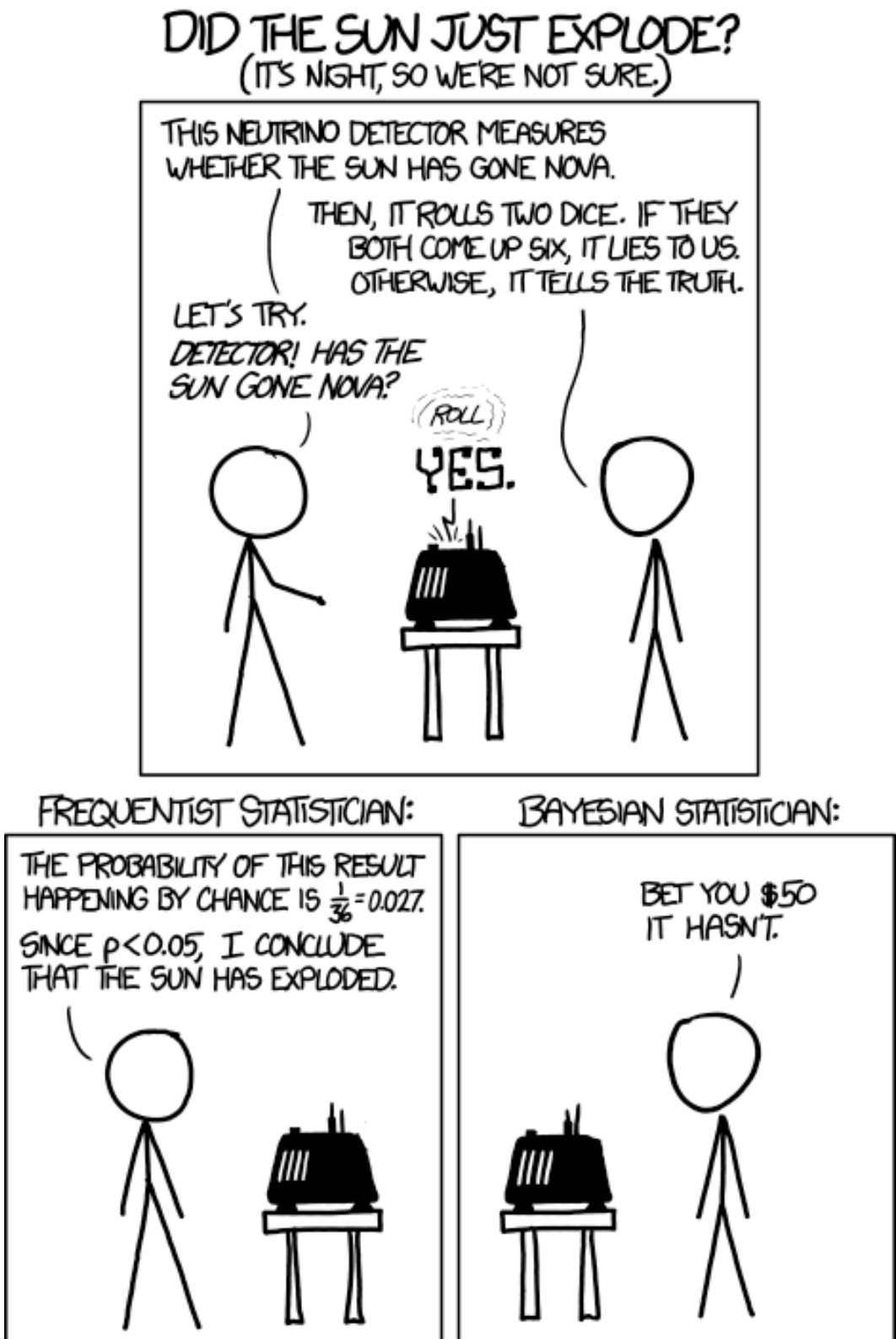
- Biological examples of Bayesian(?) information integration
- Bayesian-inspired integration of features in speech recognition
- Bayesian-inspired active listening
- Conclusions and future work

Integration of multiple modalities and information sources in biological systems

Caveat

https://www.instagram.com/p/8tcAWOkabO/

And what do I mean by “Bayesian?”



From https://xkcd.com/1132/

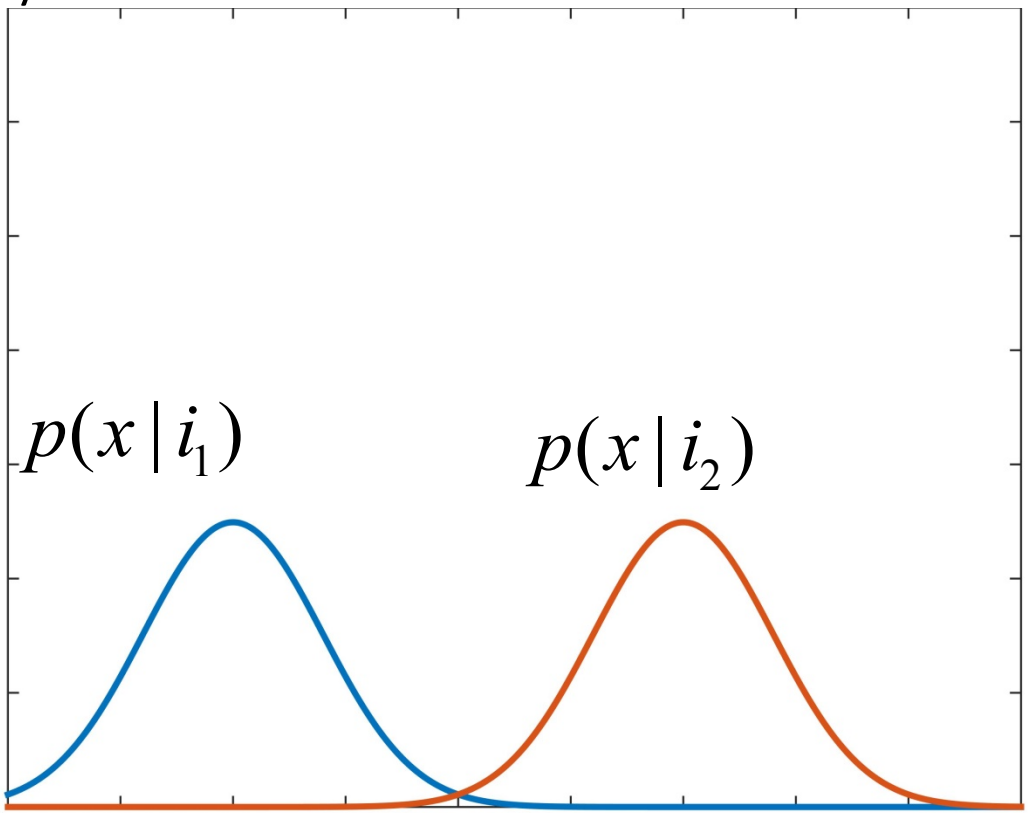
Bayesian integration of multiple information sources

In human perception, multiple modalities are integrated with prior knowledge, leading to superior performance in understanding speech and identifying environmental signals.

To understand the process of information integration, it is helpful to look at simple cases that can be tested in laboratory conditions.

Example:

Integration of visual and haptic information on height of a ridge



1. When both information sources are equally reliable:

Bayesian integration of multiple information sources

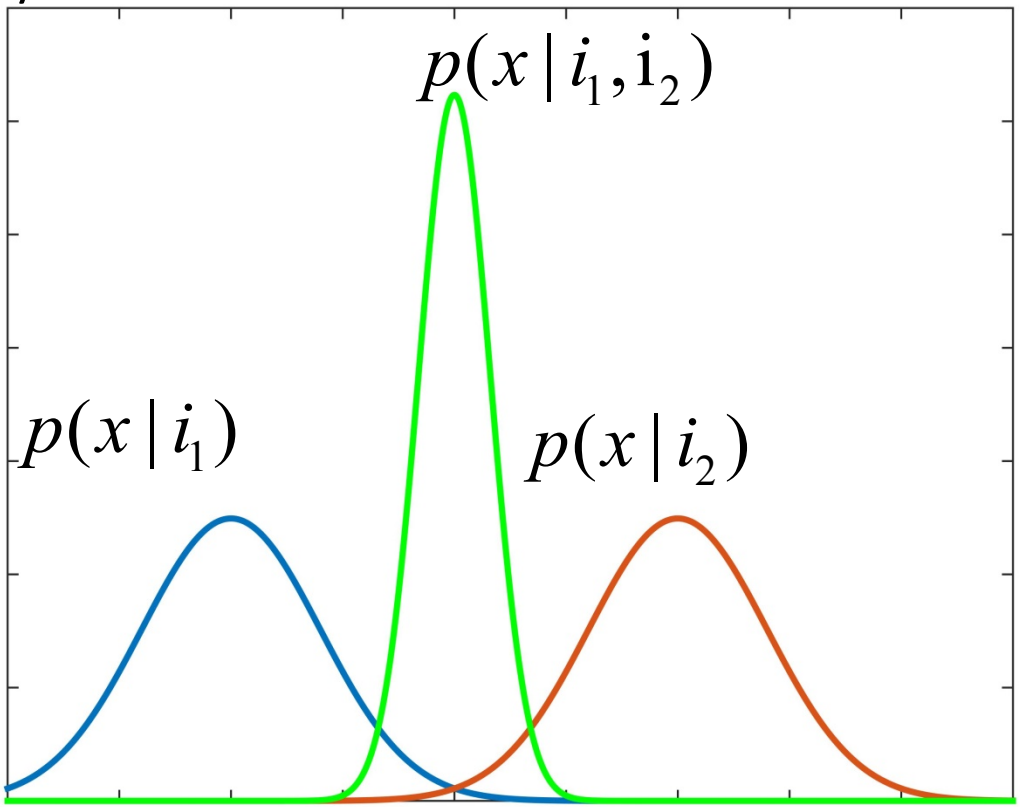
In human perception, multiple modalities are integrated with prior knowledge, leading to superior performance in understanding speech and identifying environmental signals.

To understand the process of information integration, it is helpful to look at simple cases that can be tested in laboratory conditions.

The green curve shows the posterior probability of height x, given both pieces of information.

Its expectation—the Bayes-optimal estimate in a minimum-mean-squared-error-sense—is

$$\hat{x} = \frac{\mu_1 + \mu_2}{2}$$

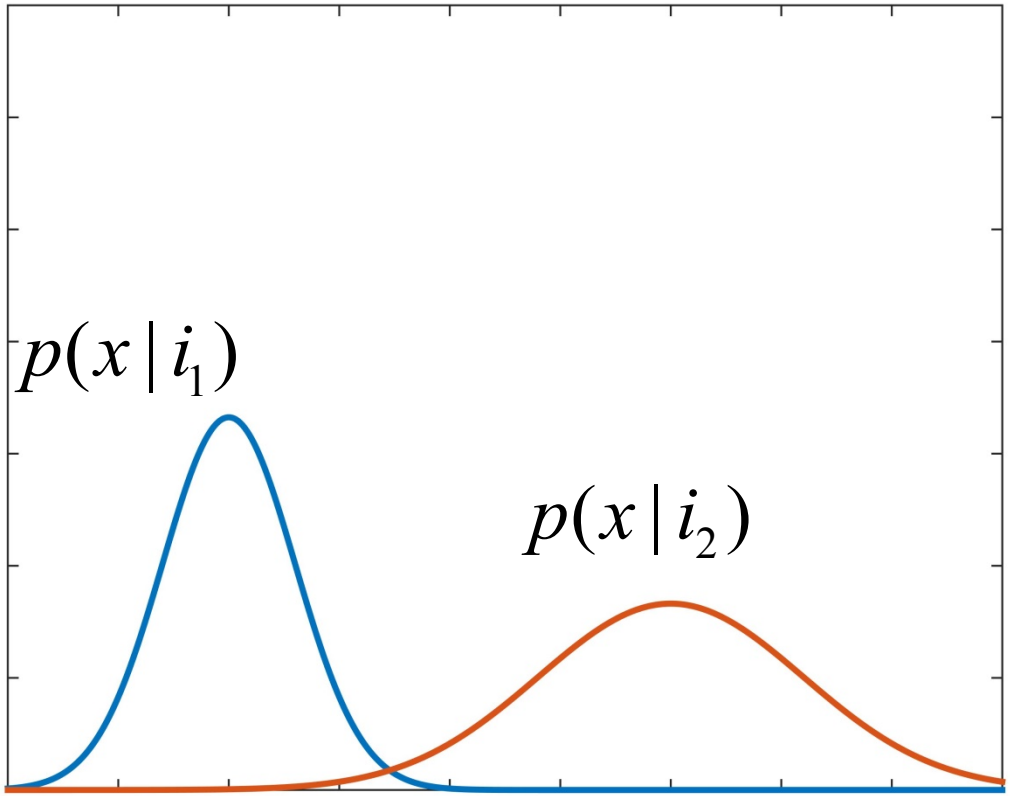


Bayesian integration of multiple information sources

In human perception, multiple modalities are integrated with prior knowledge, leading to superior performance in understanding speech and identifying environmental signals.

To understand the process of information integration, it is helpful to look at simple cases that can be tested in laboratory conditions.

but when one information source is clearly more reliable:



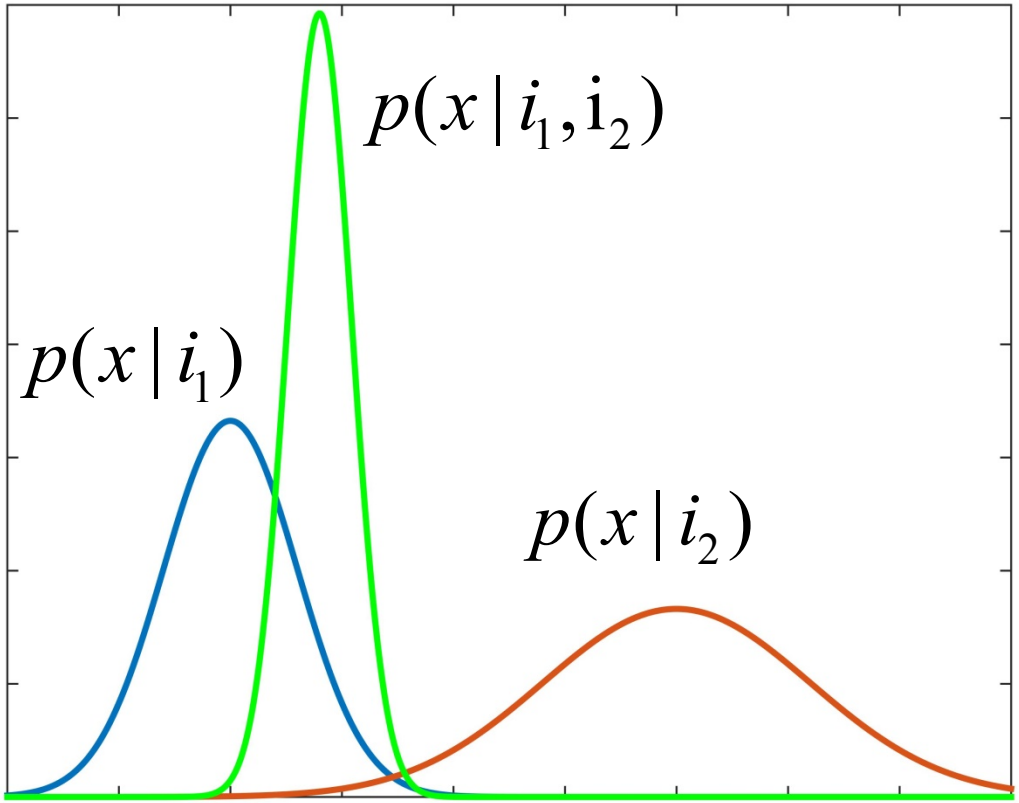
Bayesian integration of multiple information sources

In human perception, multiple modalities are integrated with prior knowledge, leading to superior performance in understanding speech and identifying environmental signals.

To understand the process of information integration, it is helpful to look at simple cases that can be tested in laboratory conditions.

the best estimate is closer to the more reliable information source:

$$\hat{x} = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$



Bayesian integration of multiple information sources

In human perception, multiple modalities are integrated with prior knowledge, leading to superior performance in understanding speech and identifying environmental signals.

To understand the process of information integration, it is helpful to look at simple cases that can be tested in laboratory conditions.

This experiment has been performed in the lab, cf. M. Ernst and M. Banks, Nature, vol. 415, 2002:

“Humans integrate visual and haptic information in a statistically optimal fashion.”

Noise was added to the visual stimuli in order to achieve varying reliability.

Bayesian integration of multiple information sources

Similar findings - quasi-Bayes-optimal integration of multiple sources of sensory information - have been reported in a number of settings, including experiments with birds and mammals, e.g.:

[Knill2004] D. Knill and A. Pouget: "The Bayesian brain: the role of uncertainty in neural coding and computation" Trends in Neuroscience, Vol. 27, No. 12, December 2004.

[Cheng2007] K. Cheng, J. Huttenlocher, S. Shettleworth and J. Rieser: "Bayesian Integration of Spatial Information," Psychological Bulletin, vol 133, no. 4, 2007.

and related observations have been made re. audio-visual recognition

[Ma2009] Wei Ji Ma, Xiang Zhou, Lars A. Ross, John J. Foxe, Lucas C. Parra: "Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space" PLoS ONE 4(3). doi:10.1371/journal.pone.0004638

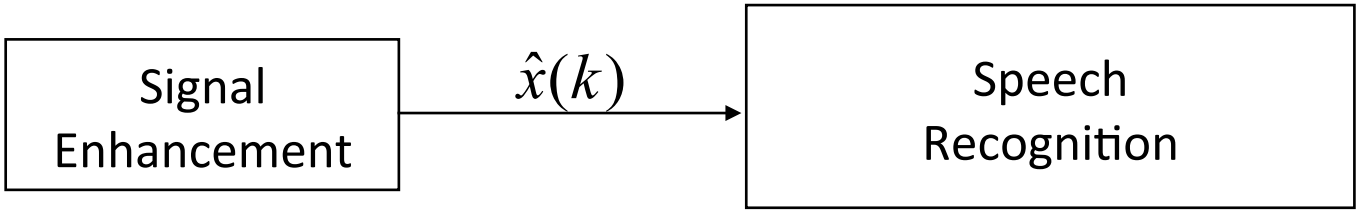
but will be addressed in next talk.

Bayesian-inspired

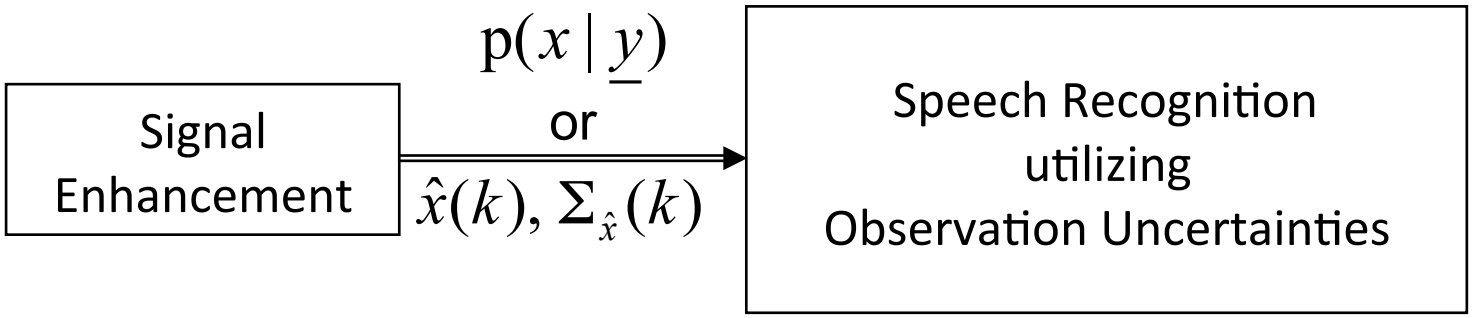
Machine Listening

Bayesian-inspired machine listening

But typical robust speech recognition systems possess this structure:



Uncertainty-of-observation strategy: Utilize statistical information where available



1 Question: How to estimate uncertainties

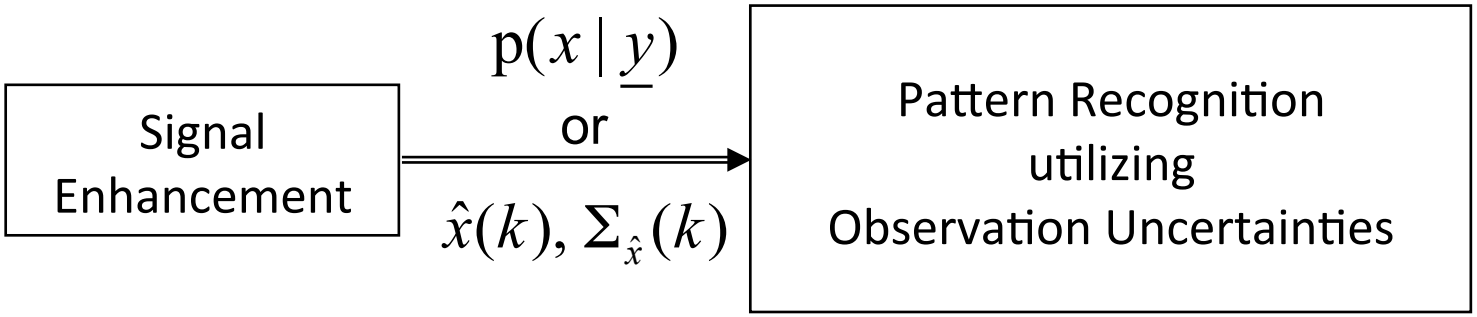
2 examples:

- Discriminative feature transform for Gaussian models
- Neural-network-based acoustic speech recognition

Bayesian-inspired Machine Listening:

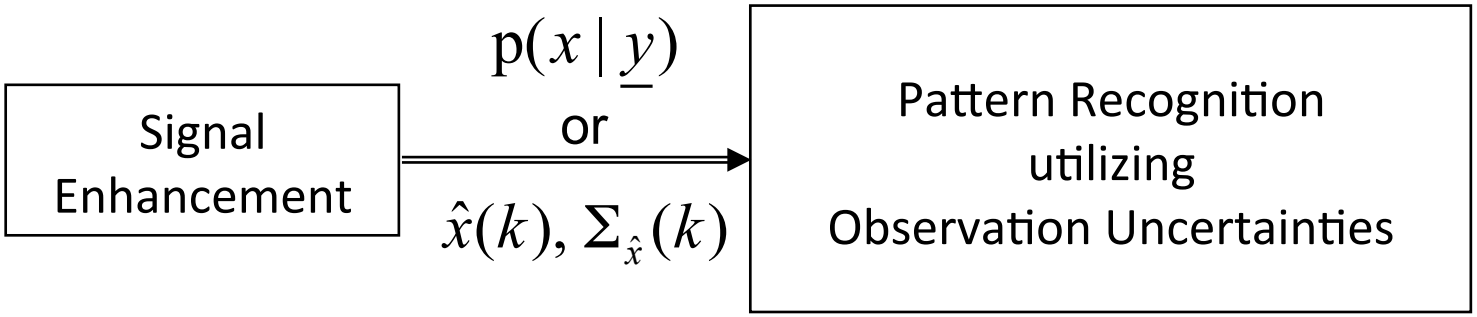
Uncertainty Estimation

Uncertainty estimation



How to estimate uncertainties?

Uncertainty estimation



How to estimate uncertainties?

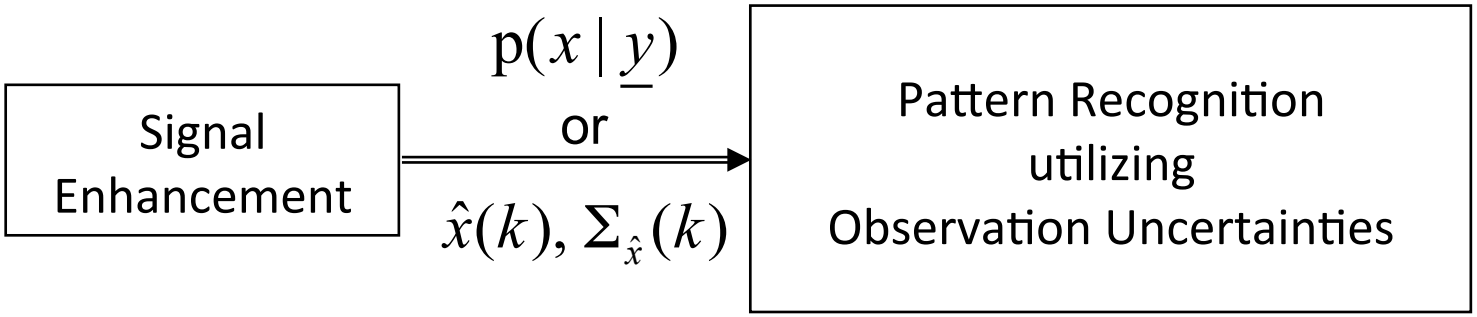
Data-driven vs. Model driven

- Learn what uncertainties/distributions look like for certain dataset or many (data-driven)
- or determine functional relationship for specific signal enhancement (model-driven)

Data-Driven, cf. [Gemmeke2010, Srinivasan2007, Tran2014, Kallasjoki2015]

Model-Driven, cf. [Astudillo2013, Nesta2013, Kolossa2005, Kolossa2010]

Uncertainty estimation

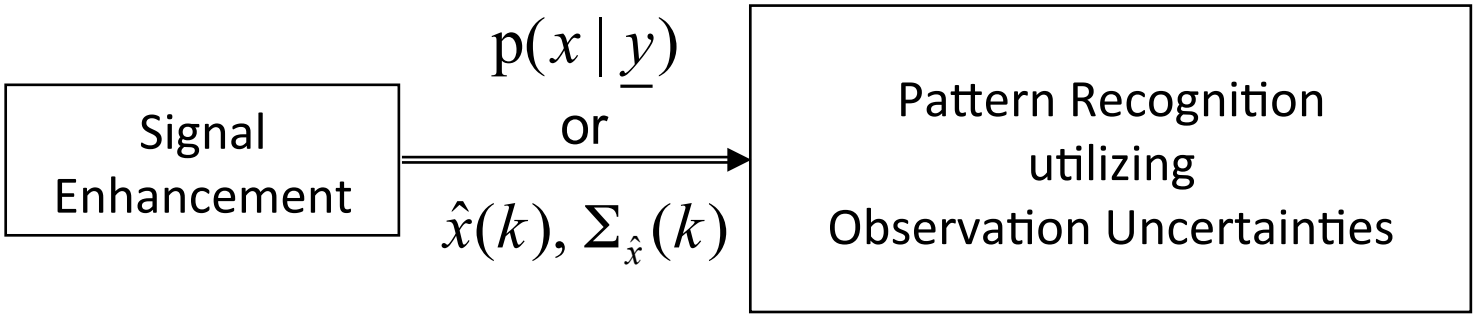


How to estimate uncertainties?

Binary vs. non-binary

- Features can be assumed to be reliable or unreliable ("glimpsing" models)
  Advantages:
  Binary reliability estimation easier than estimating higher-order parameters.
  Glimpsing seems to predict speech intelligibility in noise [Cooke2006, Barker2007]
- Features can be described by statistical model (as we do in the following)
  Advantages:
  More precise (can be Bayesian).
  Feature uncertainty can be estimated for one feature space & transformed to another [Kolossa2010, Astudillo2013].

Uncertainty estimation



How to estimate uncertainties?

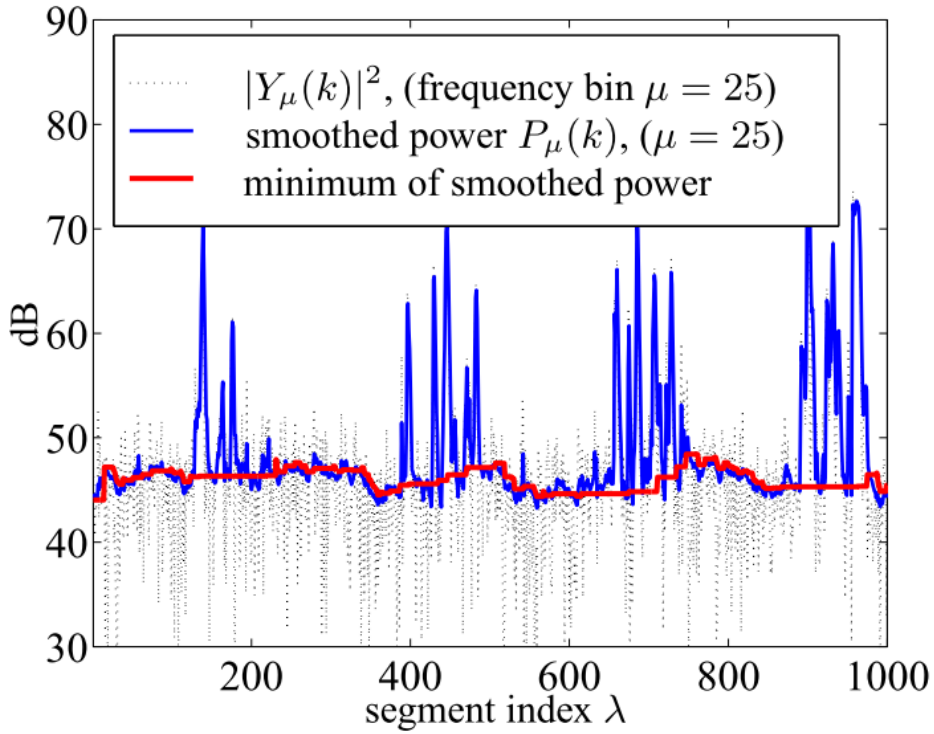
Example here

- Model-based estimation of uncertainties*, using signal-enhancement model*

Example for Single Channel Case

Improved Minima-Controlled Recursive Averaging (IMCRA)

Estimate and track noise by observing that spectral minima are due to noise alone:

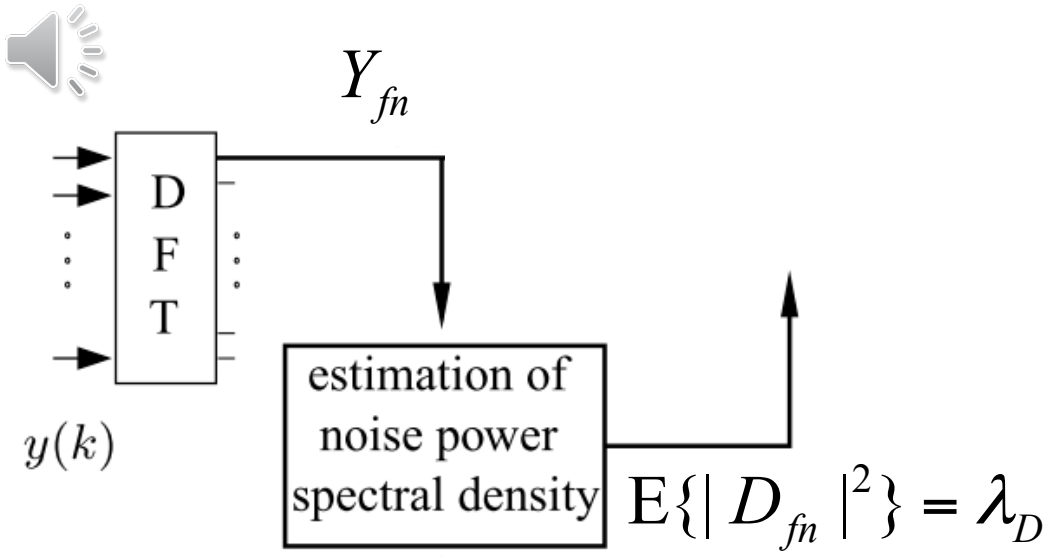


from R. Martin: “Statistical methods for the enhancement of noisy speech,” Proc. IWAENC 2003, see also [Martin2001, Cohen2003].

Example for Single Channel Case

Improved Minima-Controlled Recursive Averaging (IMCRA)

Estimate and track noise in recursive system:



based on R. Martin: “Statistical methods for the enhancement of noisy speech,” Proc. IWAENC 2003.

Example for Single Channel Case

Speech Estimation

Minimize residual error

$$\hat{X}_{fn}^{W} = \mathrm{E}\{X_{fn} \mid Y\}$$

with IMCRA-estimated noise variance λ_D and estimated speech variance λ_X

Uncertainty Estimation

Use residual error of MMSE-Estimator

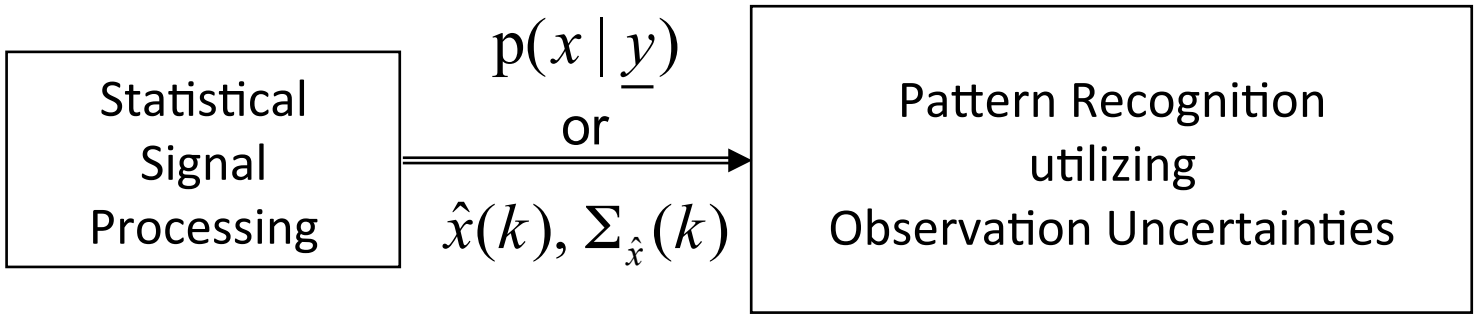
$$\mathrm{E}\{\| X_{fn} - \hat{X}_{fn}^{W} \|^2 \mid Y\} = \frac{\lambda_X \lambda_D}{\lambda_X + \lambda_D}$$

[Astudillo2013] Ramón Fernández Astudillo and Reinhold Orglmeister: “Computing MMSE Estimates and Residual Uncertainty Directly in the Feature Domain of ASR using STFT Domain Speech Distortion Models” IEEE Trans. Audio Speech and Language Processing, Vol. 21, No. 5, May 2013.

Bayesian-inspired Machine Listening:

Rapid Adaptation

Rapid adaptation in speech recognition

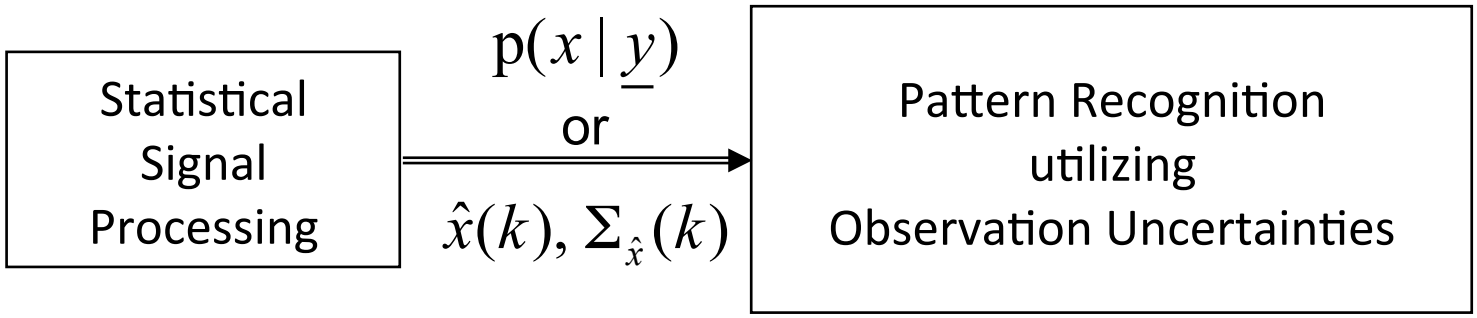


Approaches to observation uncertainty using HMM/GMM-Systems

- ***Front-End-Techniques***: (Bounded) Marginalization, Imputation, Uncertainty Decoding, Predictive Decoding, Modified Imputation, Significance Decoding [Cooke2001, Raj2005, Deng2005, Kolossa2005, Ion2008, Astudillo2013, Abdelaziz2013]

  -> modify likelihood computation

Rapid adaptation in speech recognition



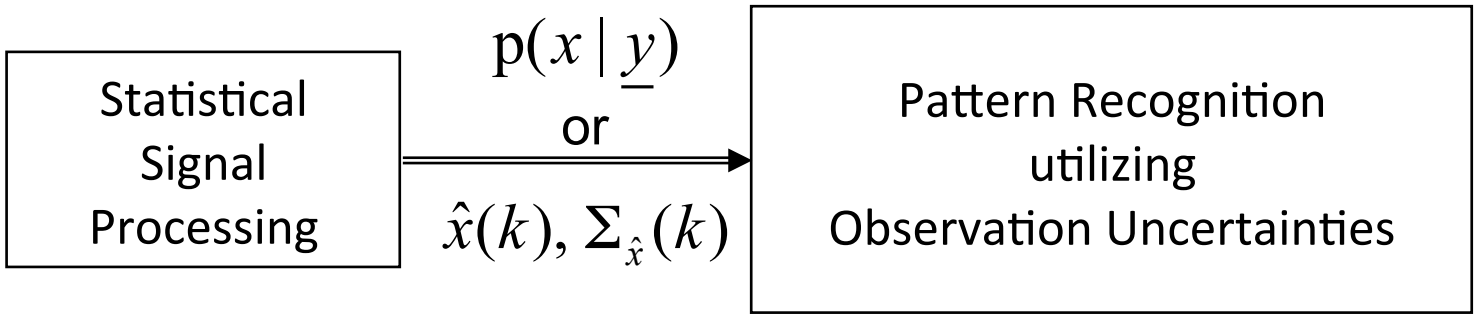
Approaches to observation uncertainty using HMM/GMM-Systems

- ***Front-End-Techniques***: (Bounded) Marginalization, Imputation, Uncertainty Decoding, Predictive Decoding, Modified Imputation, Significance Decoding [Cooke2001, Raj2005, Deng2005, Kolossa2005, Ion2008, Astudillo2013, Abdelaziz2013]

  -> modify likelihood computation

Binary uncertainty

Rapid adaptation in speech recognition



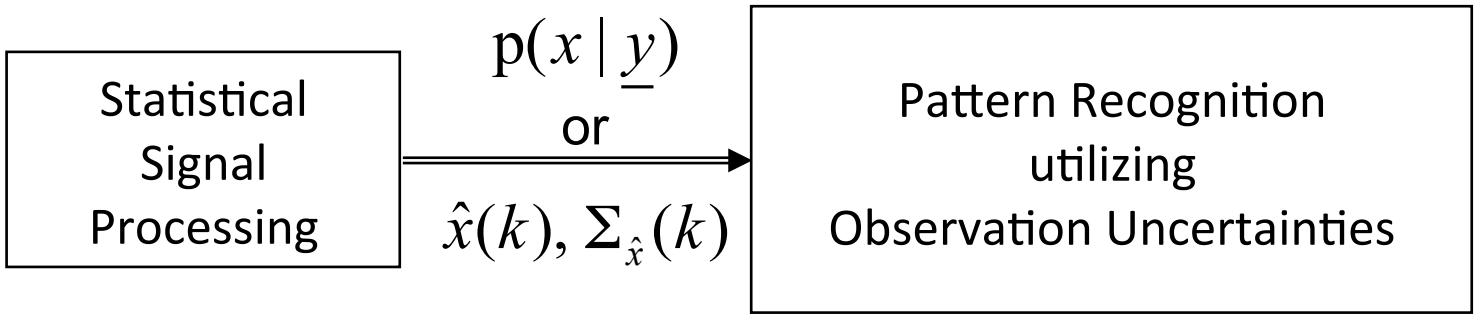
Approaches to observation uncertainty using HMM/GMM-Systems

- ***Front-End-Techniques***: (Bounded) Marginalization, Imputation, Uncertainty Decoding, Predictive Decoding, Modified Imputation, Significance Decoding [Cooke2001, Raj2005, Deng2005, Kolossa2005, Ion2008, Astudillo2013, Abdelaziz2013]

  -> modify likelihood computation

Continuous uncertainty (variance)

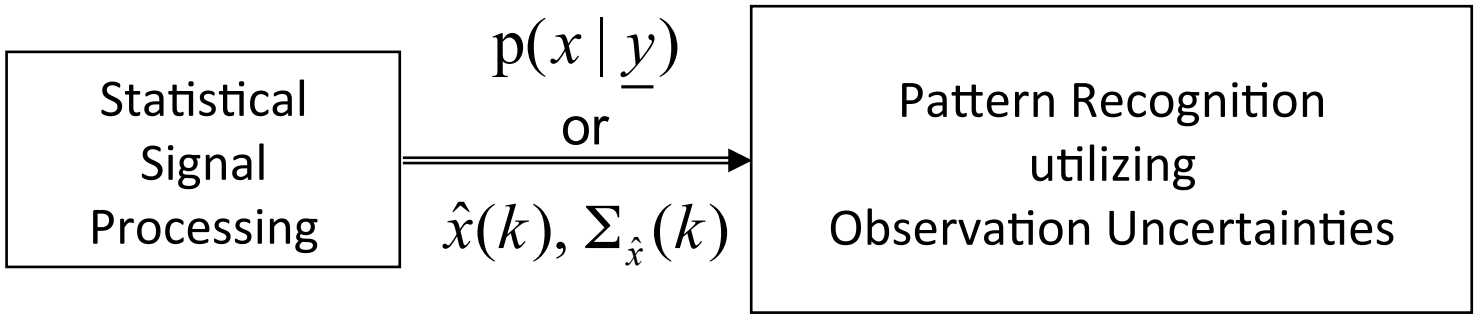
Rapid adaptation in speech recognition



Approaches to observation uncertainty using HMM/GMM-Systems

- ***Front-End-Techniques***: (Bounded) Marginalization, Imputation, Uncertainty Decoding, Predictive Decoding, Modified Imputation, Significance Decoding [Cooke2001, Raj2005, Deng2005, Kolossa2005, Ion2008, Astudillo2013, Abdelaziz2013]
  -> modify likelihood computation
- ***Feature transform techniques:*** Noise-adaptive LDA [Kolossa2013]
  -> transform features (and compute model update)
- ***Back-end-technique***: Joint uncertainty decoding [Liao2005]
  -> adapts model (similar to constrained MLLR + variance bias)

Rapid adaptation in speech recognition



Approaches to observation uncertainty using HMM/GMM-Systems

- ***Front-End-Techniques***: (Bounded) Marginalization, Imputation, Uncertainty Decoding, Predictive Decoding, Modified Imputation, Significance Decoding [Cooke2001, Raj2005, Deng2005, Kolossa2005, Ion2008, Astudillo2013, Abdelaziz2013]
  -> modify likelihood computation
- ***Feature transform techniques:*** **Noise-adaptive LDA [Kolossa2013]**
  -> transform features (and compute model update)
- ***Back-end-technique***: Joint uncertainty decoding [Liao2005]
  -> adapts model (similar to constrained MLLR + variance bias)

Noise-adaptive LDA

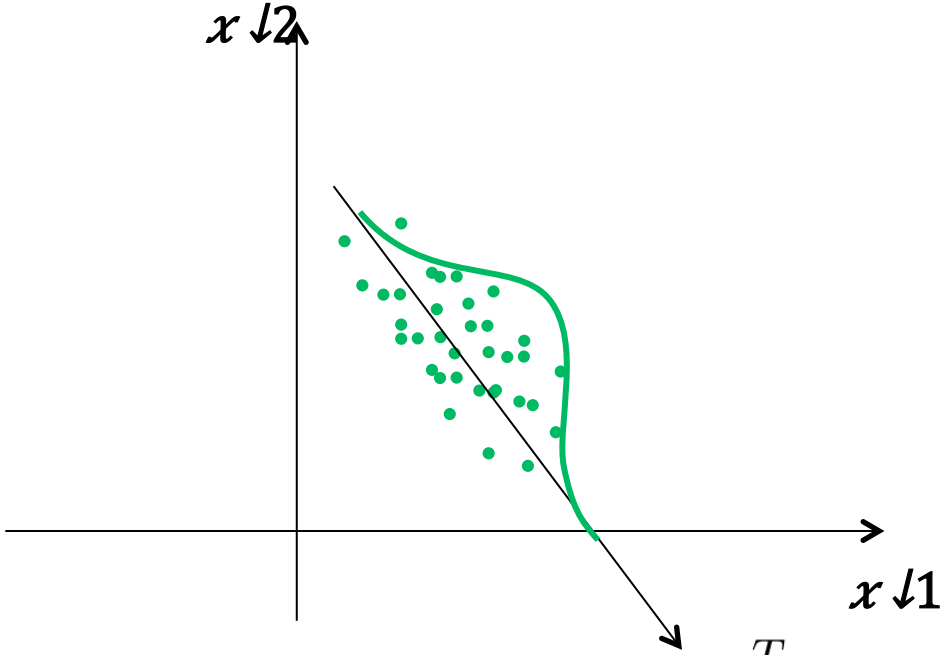
Core idea

At any point in time: Rely most on those dimensions of feature vector that are most reliable

For this purpose: Project feature vector onto lower-dimensional subspace, maximizing discriminability of projected data

What projection should we use?

Projection 1: Principal Components Analysis

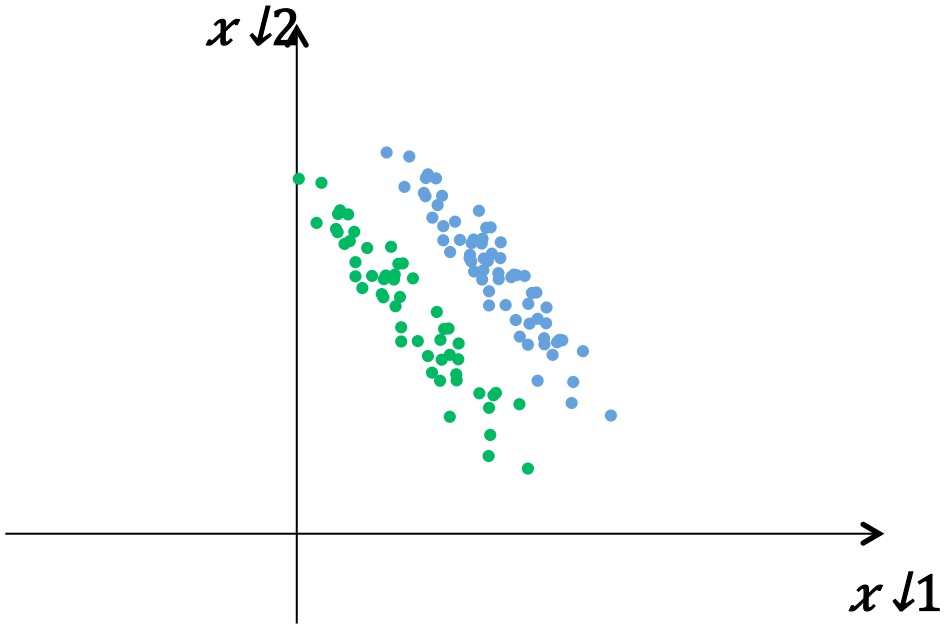


$$\hat{x}'_1 = \mathbf{w}_{1,\mathrm{PCA}}^T \hat{\mathbf{x}} \text{ with } \hat{\mathbf{x}} = [\hat{x}_1, \hat{x}_2]^T$$

For projection onto component i, PCA uses the i^{th} largest eigenvector of the total covariance matrix: $\boldsymbol{\Sigma}_t \mathbf{w}_{i,\mathrm{PCA}} = \lambda_i \mathbf{w}_{i,\mathrm{PCA}}$

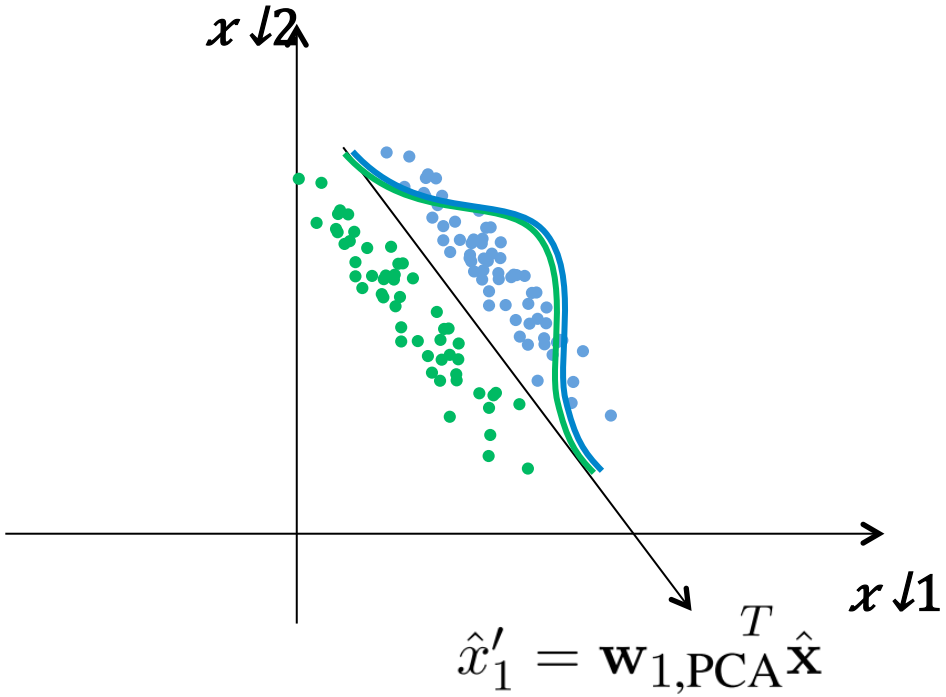
Projection 1: Principal Components Analysis

Minor problem when using PCA in classification:



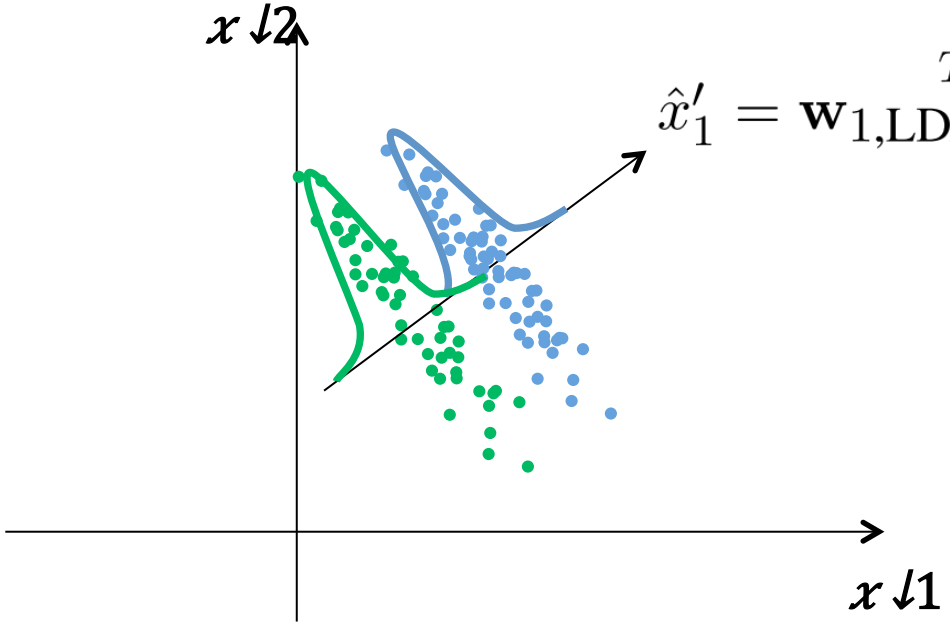
Projection 1: Principal Components Analysis

Minor problem when using PCA in classification:



“Adidas problem”

Projection 2: Linear Discriminant Analysis

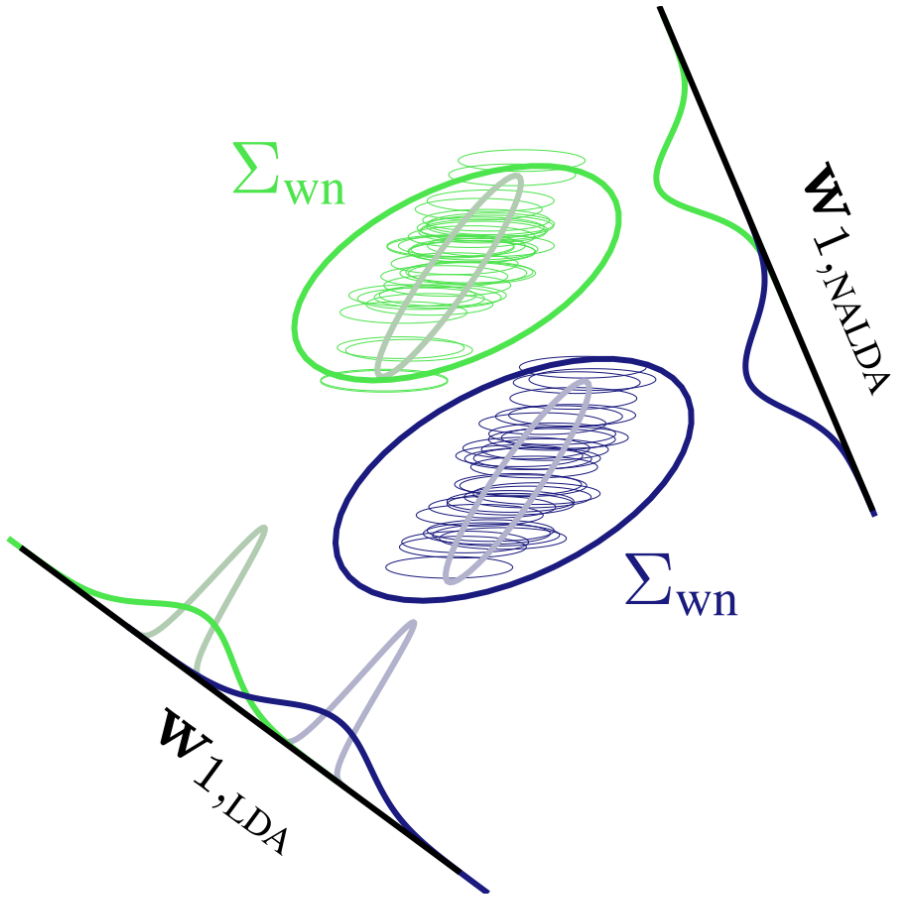


$$\hat{x}'_1 = \mathbf{w}_{1,\mathrm{LDA}}^T \hat{\mathbf{x}} \quad \text{with} \quad \hat{\mathbf{x}} = [\hat{x}_1, \hat{x}_2]^T$$

For projection onto component i, LDA uses the i^{th} largest generalized eigenvector of between-class covariance $\mathbf{\Sigma}_b$ and within-class covariance $\mathbf{\Sigma}_w$:

$$\mathbf{\Sigma}_b \, \mathbf{w}_{i,\mathrm{LDA}} = \lambda_i \, \mathbf{\Sigma}_w \, \mathbf{w}_{i,\mathrm{LDA}}$$

Projection 3: *Noise-Adaptive* Linear Discriminant Analysis (NALDA)



LDA vs. NALDA on uncertain features

Noise-Adaptive Linear Discriminant Analysis

Results on 2nd CHiME challenge dataset [Vincent2013]

	Keyword Error Rates (%) on CHiME 2 Corpus						
SNR	-6dB	-3dB	0dB	3dB	6dB	9dB	avg.
HTK (matched) Baseline	50.7	41.3	32.5	24.9	21.2	17.1	31.3

[Kolossa2013] D. Kolossa, S. Zeiler, R. Saeidi, R.F. Astudillo: "Noise-Adaptive LDA: A New Approach for Speech Recognition Under Observation Uncertainty, IEEE Signal Processing Letters 20 (11), pp. 1018-1021, 2013.

Noise-Adaptive Linear Discriminant Analysis

Results on 2nd CHiME challenge dataset [Vincent2013]

	Keyword Error Rates (%) on CHiME 2 Corpus						
SNR	-6dB	-3dB	0dB	3dB	6dB	9dB	avg.
HTK (matched) Baseline	50.7	41.3	32.5	24.9	21.2	17.1	31.3
Using Oracle Uncertainties							
Diagonal Gaussian Mixture Model	27.9	23.0	18.1	15.4	12.8	10.4	17.9
NALDA	17.3	13.2	11.5	9.3	7.7	7.6	11.1

[Kolossa2013] D. Kolossa, S. Zeiler, R. Saeidi, R.F. Astudillo: "Noise-Adaptive LDA: A New Approach for Speech Recognition Under Observation Uncertainty, IEEE Signal Processing Letters 20 (11), pp. 1018-1021, 2013.

Noise-Adaptive Linear Discriminant Analysis

Results on 2nd CHiME challenge dataset [Vincent2013]

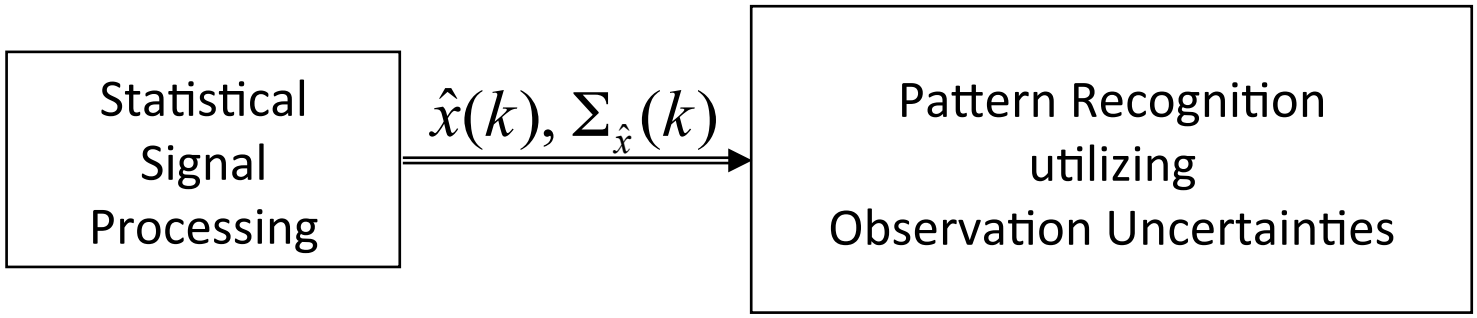
	Keyword Error Rates (%) on CHiME 2 Corpus						
SNR	-6dB	-3dB	0dB	3dB	6dB	9dB	avg.
HTK (matched) Baseline	50.7	41.3	32.5	24.9	21.2	17.1	31.3
Using Oracle Uncertainties							
Diagonal Gaussian Mixture Model	27.9	23.0	18.1	15.4	12.8	10.4	17.9
NALDA	17.3	13.2	11.5	9.3	7.7	7.6	11.1
Using Estimated Uncertainties							
Diagonal Gaussian Mixture Model	28.1	20.9	17.4	12.2	8.4	8.4	15.9
NALDA	26.0	21.1	14.8	9.1	7.6	6.7	14.2

[Kolossa2013] D. Kolossa, S. Zeiler, R. Saeidi, R.F. Astudillo: "Noise-Adaptive LDA: A New Approach for Speech Recognition Under Observation Uncertainty, IEEE Signal Processing Letters 20 (11), pp. 1018-1021, 2013.

Bayesian-inspired Machine Speech "Perception":

Processing Uncertainties in DNN/HMMs

Approaches to handling uncertain data



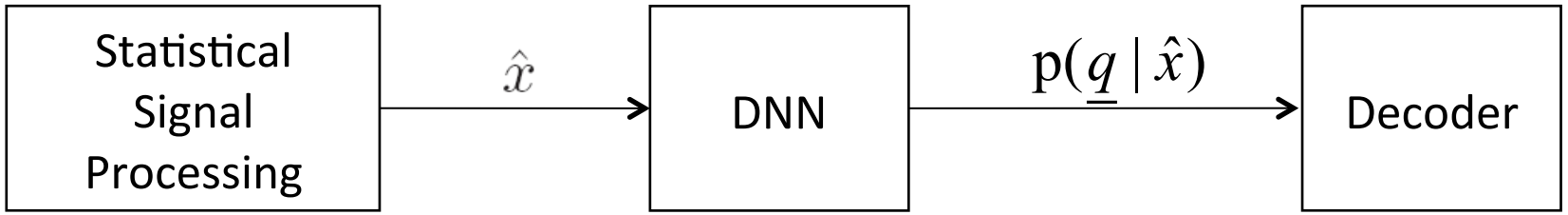
2 variants of handling uncertain observations:

- Noise-adaptive LDA
- Modified DNN decoding

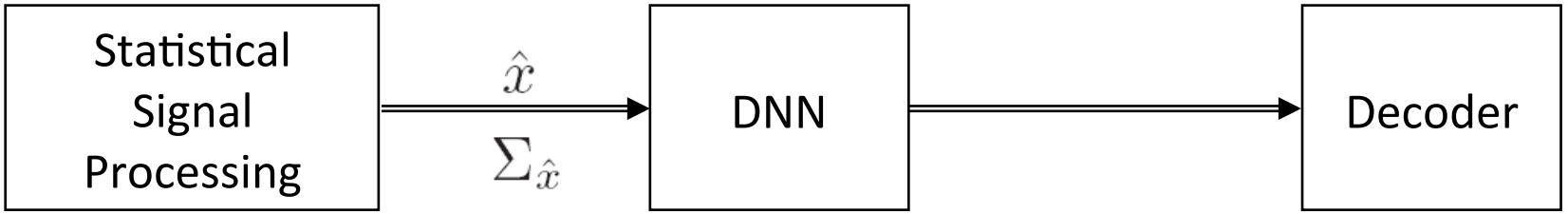
Example 2: Neural-network-based speech recognition

But how to use observation uncertainties in neural-network-based speech recognition?

Architecture of Hybrid Deep-Neural-Network (DNN)-based ASR:



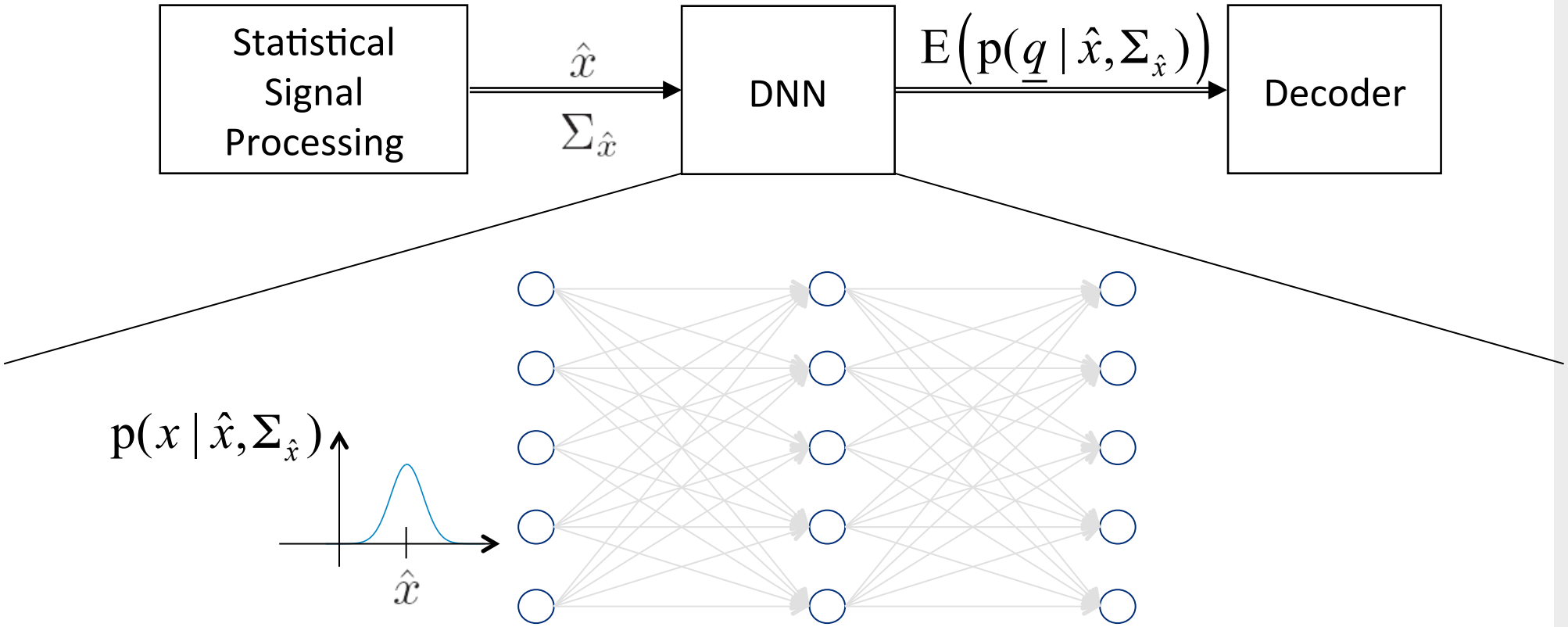
+ Observation Uncertainties



2 solution principles:

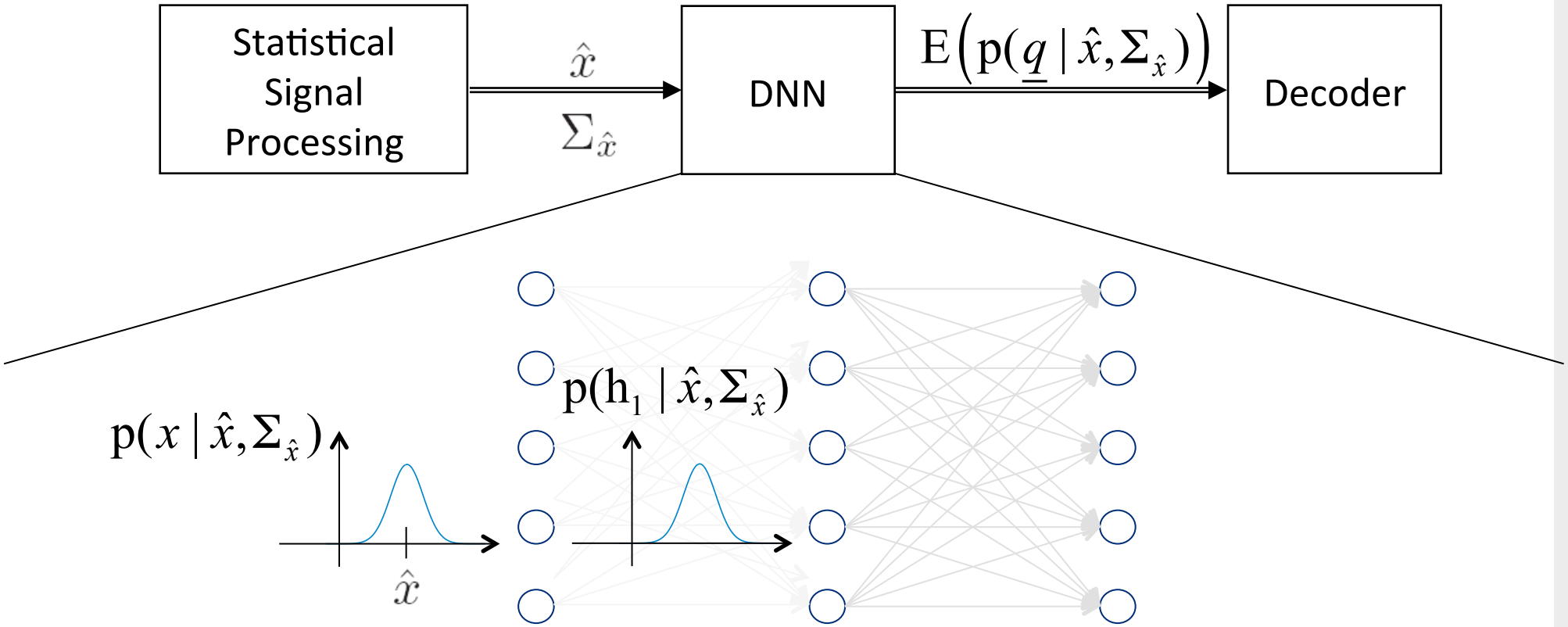
Example 2: Neural-network-based speech recognition

Solution 1: Use Monte-Carlo approximation of probability distribution throughout network



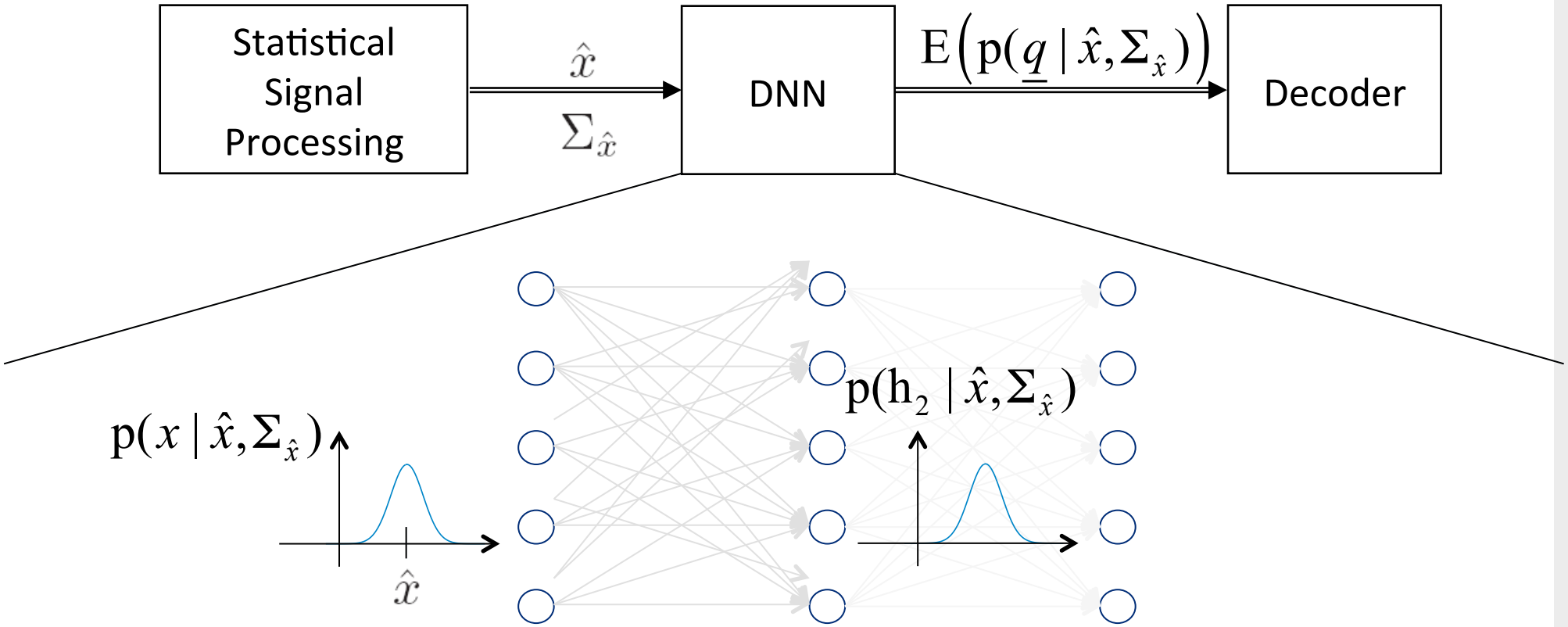
Example 2: Neural-network-based speech recognition

Solution 1: Use Monte-Carlo approximation of probability distribution throughout network



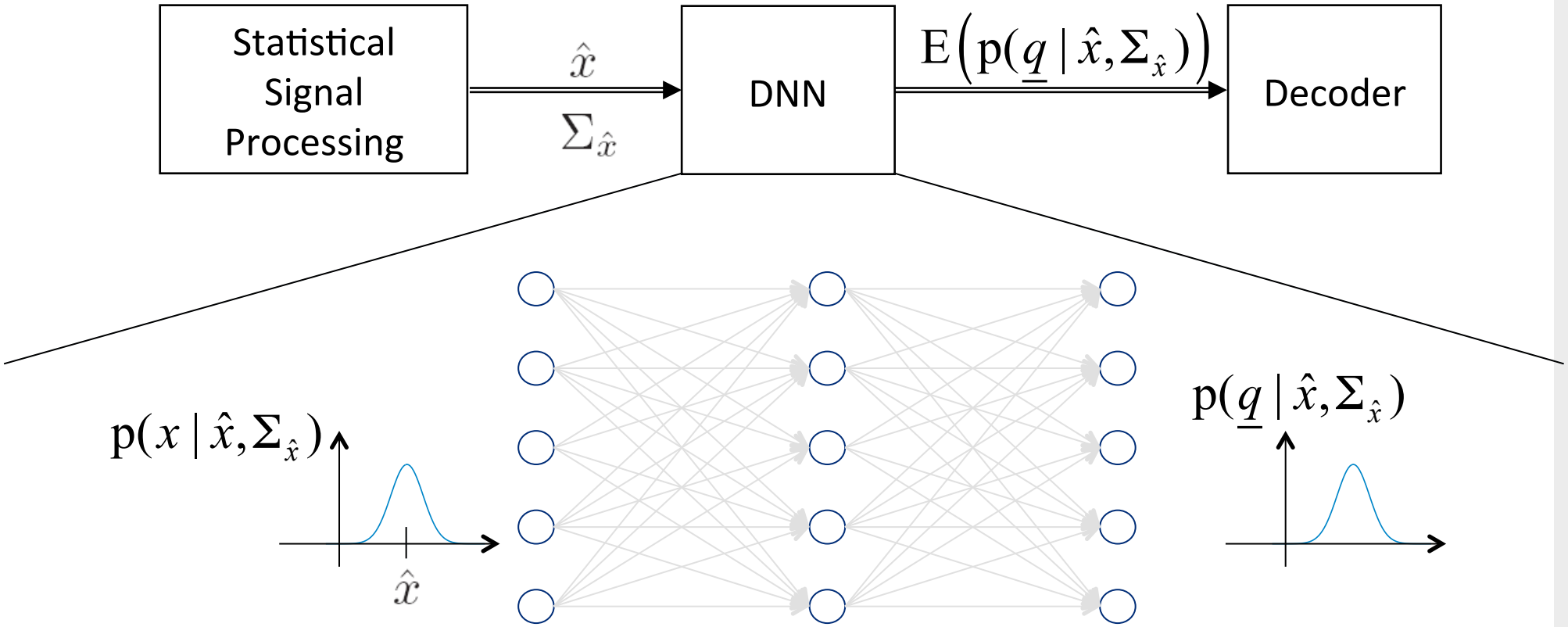
Example 2: Neural-network-based speech recognition

Solution 1: Use Monte-Carlo approximation of probability distribution throughout network



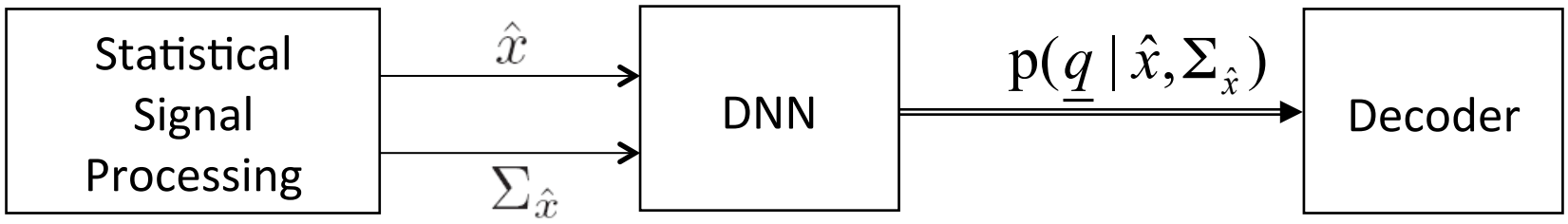
Example 2: Neural-network-based speech recognition

Solution 1: Use Monte-Carlo approximation of probability distribution throughout network



Example 2: Neural-network-based speech recognition

Solution 2: Use observation uncertainties as additional features



Use uncertainties in DNN training ("Supervised")

-> How do both versions perform in a medium-vocabulary robust speech recognition task?

[Abdelaziz2015] A. Hussen Abdelaziz, Shinji Watanabe, John R. Hershey, Emanuel Vincent, Dorothea Kolossa : "On Decoding of Uncertain Data Using Deep Neural Networks," Proc. Interspeech 2015

Example 2: Neural-network-based speech recognition

Experimental Setup

Wall-Street Journal task (5,000 words) of second CHiME challenge [13]

Noisy data created by convolving clean utterances with binaural room impulse responses (BRIRs) and adding background noise signals at
six different SNRs: -6, -3, 0, 3, 6, and 9 dB, recorded in a domestic living room.

Baseline

[Tachioka2013] Y. Tachioka, S. Watanabe, J. L. Roux, J. R. Hershey: “Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark,” in The 2nd International Workshop on Machine Listening in Multisource Environments (CHiME), Vancouver, Canada, 2013.

	Dev. Set
Best 2	27.61
(+DLM)	27.14
(+MBR)	27.10
(+both)	26.86

Example 2: Neural-network-based speech recognition

Comparing both observation uncertainty approaches

Monte-Carlo: No network adaptation, probabilistic input, uncertainty propagation.

Uncertainty Features: Re-training of network using uncertainties as additional features.

	Dev. Set	Test Set
Baseline	27.59	21.67
Monte-Carlo (≈ Bayesian)	27.03	21.06
Uncertainty Features	26.49	20.33

[Abdelaziz2015] Ahmed Hussen Abdelaziz, Shinji Watanabe, John R. Hershey, Emanuel Vincent, Dorothea Kolossa : "On Decoding of Uncertain Data Using Deep Neural Networks," Proc. Interspeech 2015.

-> Further understanding of the best encoding of uncertain information in artificial neural networks is needed!

Bayesian-Inspired

Active perception

Active Perception

TWO!EARS-Project

Active, integrated perception in robot and machine audition

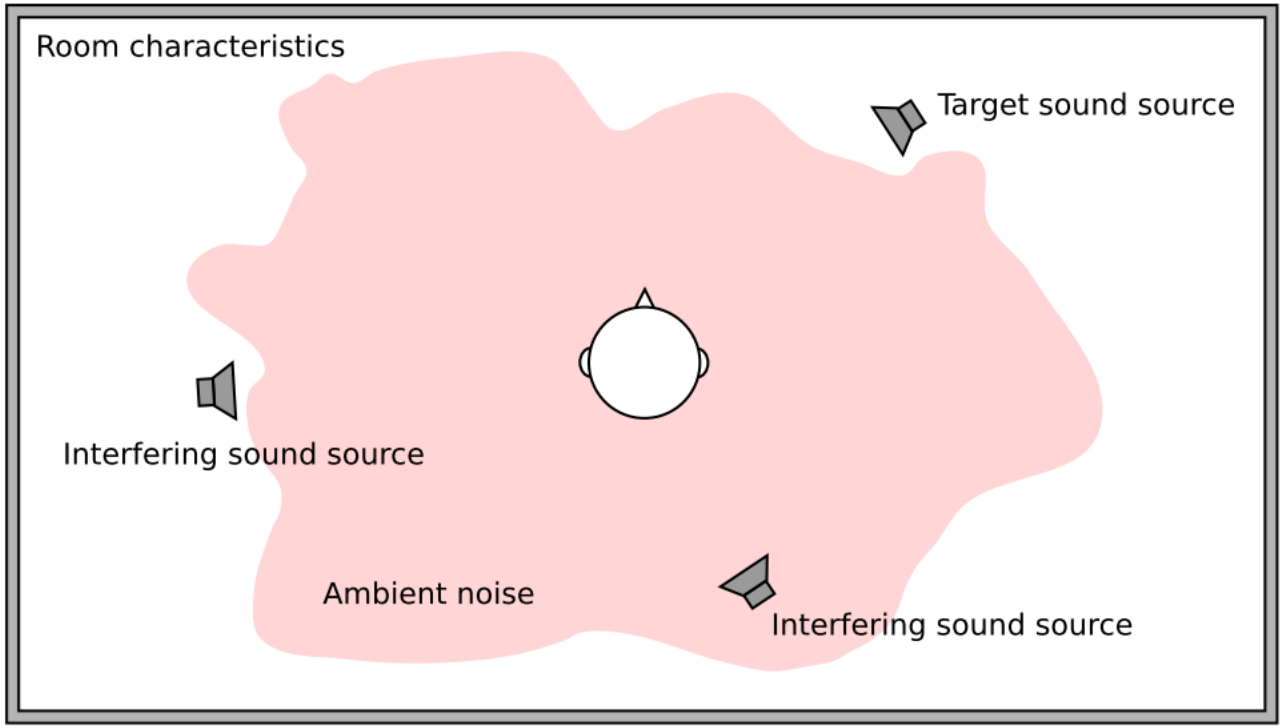
Partners:
TU Ilmenau, DTU Copenhagen, Ruhr-Universität Bochum, LAAS-CNRS Toulouse, TU Berlin, TU Eindhoven, UPMC Paris, Universität Rostock, University of Sheffield, Rensselaer Polytechnic Institute

Goals:
Understand and implement active, exploratory listening, environmental scene understanding and quality evaluation

Active Perception

A Bayesian view

Active perception is the process of incrementally refining the understanding of one's current environment through appropriate actions.

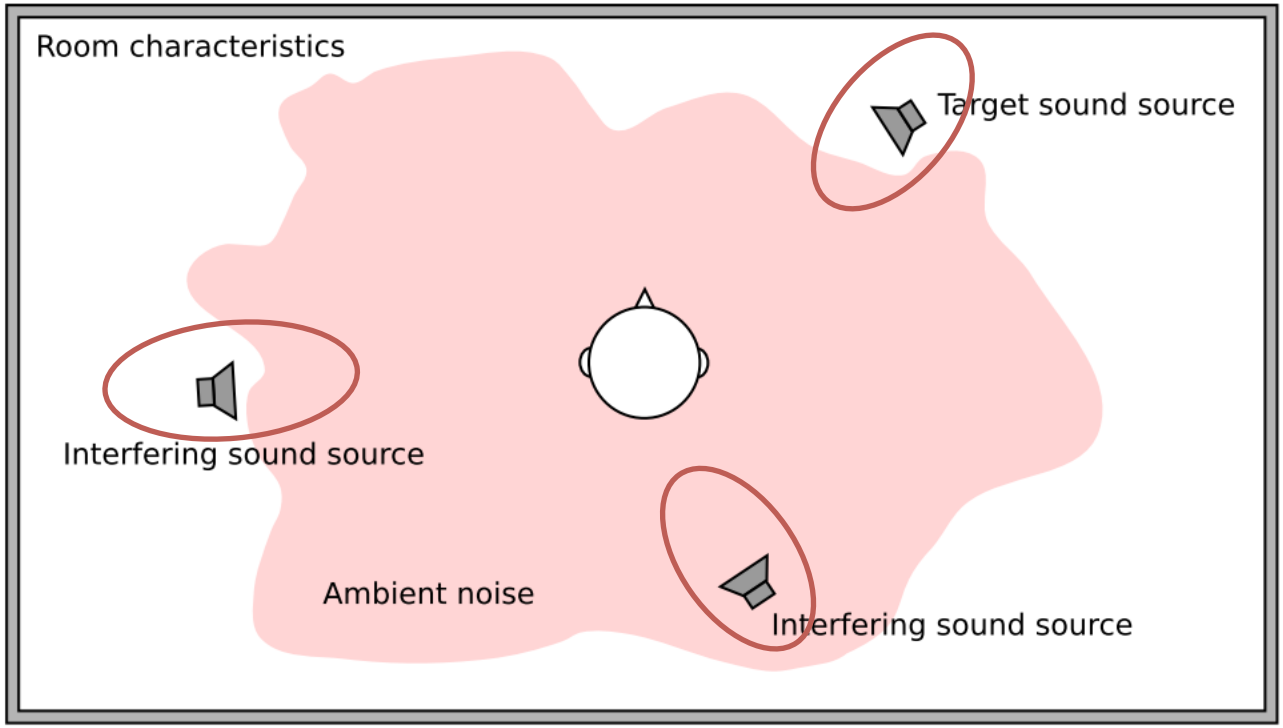


[Schymura2014] Ch. Schymura C. Schymura, T. Walther, D. Kolossa, N. Ma and G. Brown: "Binaural Sound Source Localisation using a Bayesian-network-based Blackboard System and Hypothesis-driven Feedback," Proc. Forum Acusticum, Krakow, September 2014.

Active Perception

A Bayesian view

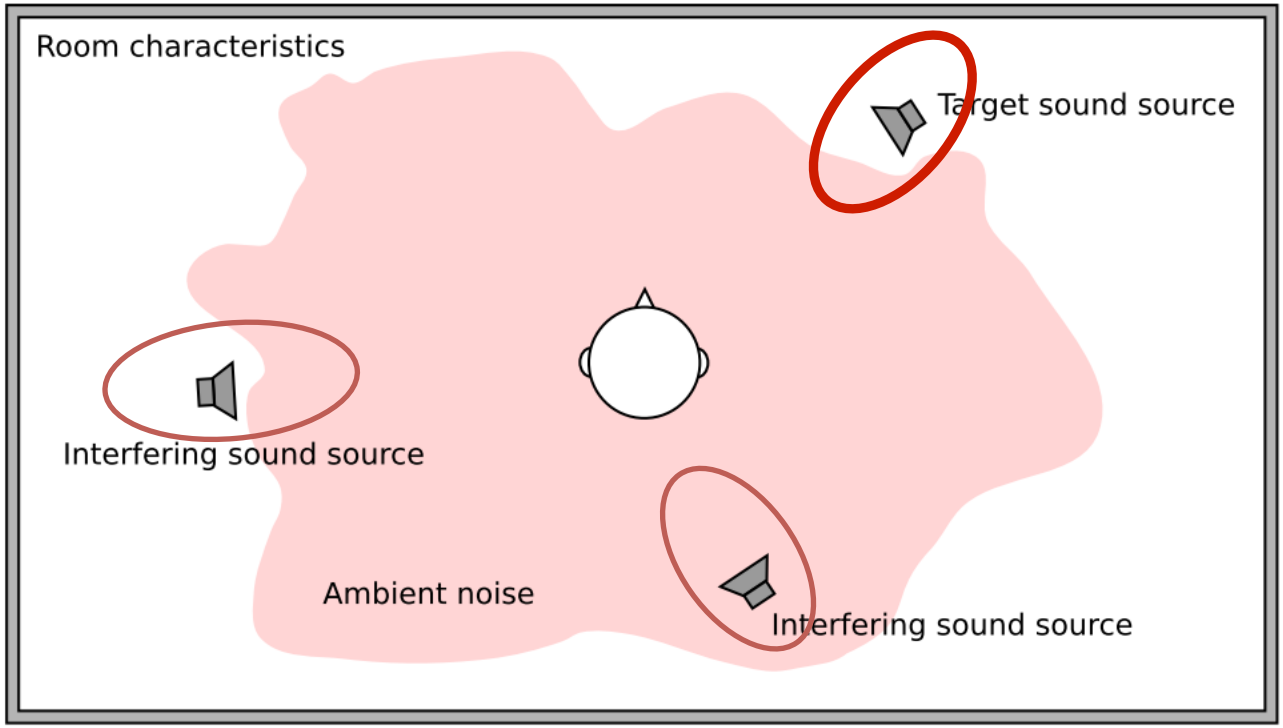
The understanding of the environment is reflected through a probabilistic description of the type and location of all present sources (the 'world model.')



Active Perception

A Bayesian view

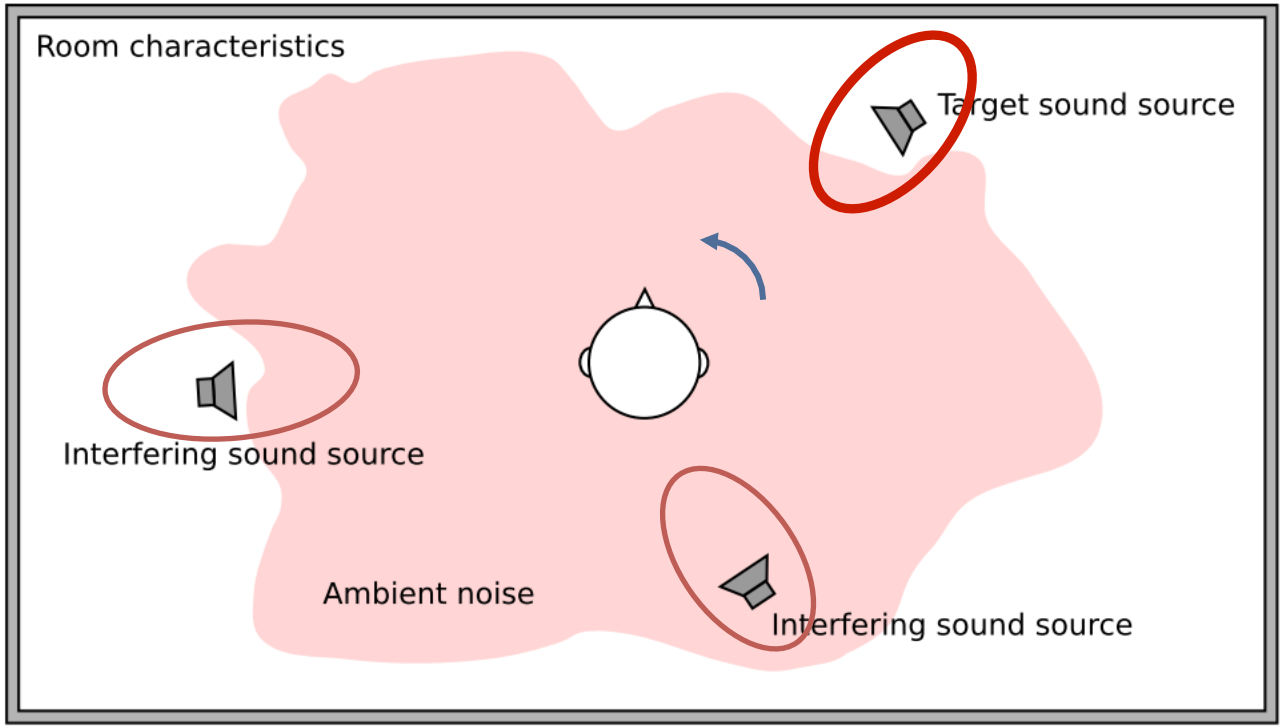
Our goal is the optimal planning of movement, in such a way as to minimize the uncertainty, focusing first on the most relevant sources.



Active Perception

A Bayesian view

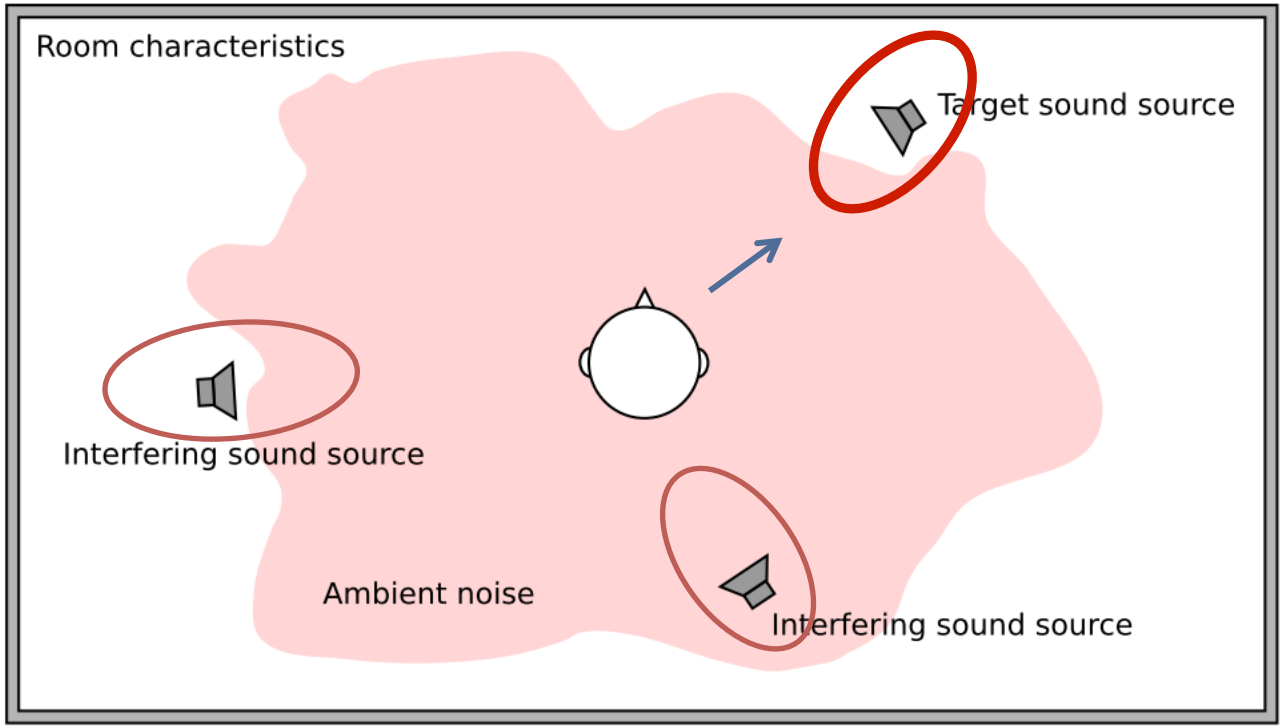
*Movement planning concerns both **head rotations**...*



Active Perception

A Bayesian view

*...and **translatory robot movements*** (addressed originally by [Ferreira2013]).

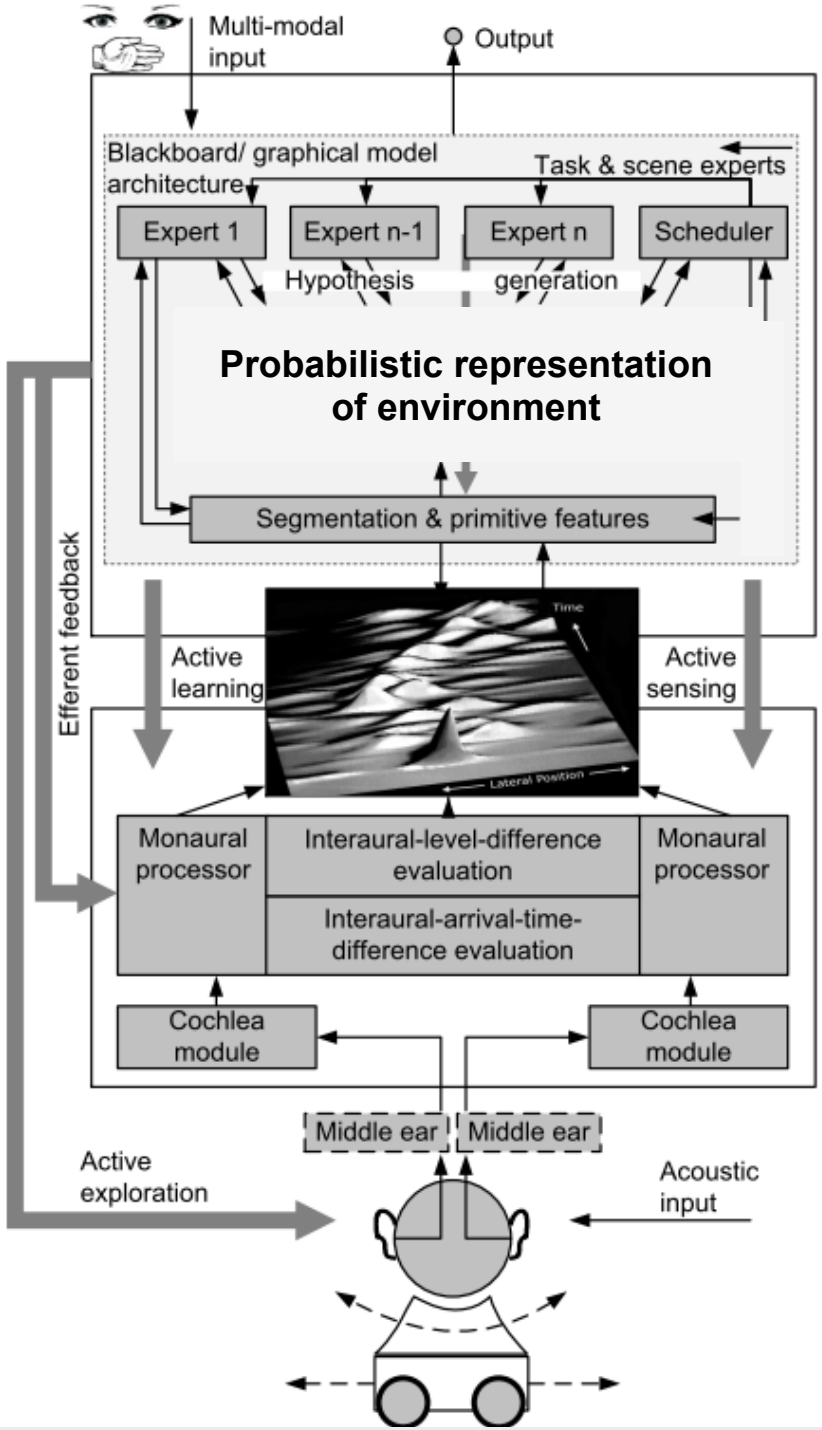


[Ferreira2013] J. Ferreira, J. Lobo, P. Bessiere, M. Castelo-Branco, and J. Dias: “A Bayesian Framework for Active Artificial Perception,” IEEE Trans. Cybernetics, vol. 43, Apr. 2013.

Active Perception

Approach:
Iteratively construct a comprehensive, probabilistic understanding of the machine in its environment, integrating bottom-up and top-down processing.

From this evolving understanding, derive best next action.



[Blauert2013] J. Blauert, D. Kolossa, K. Obermayer, and K. Adiloglu: "Further challenges and the road ahead", in J. Blauert (ed.) The technology of binaural listening, Springer, 2013.

Active Perception

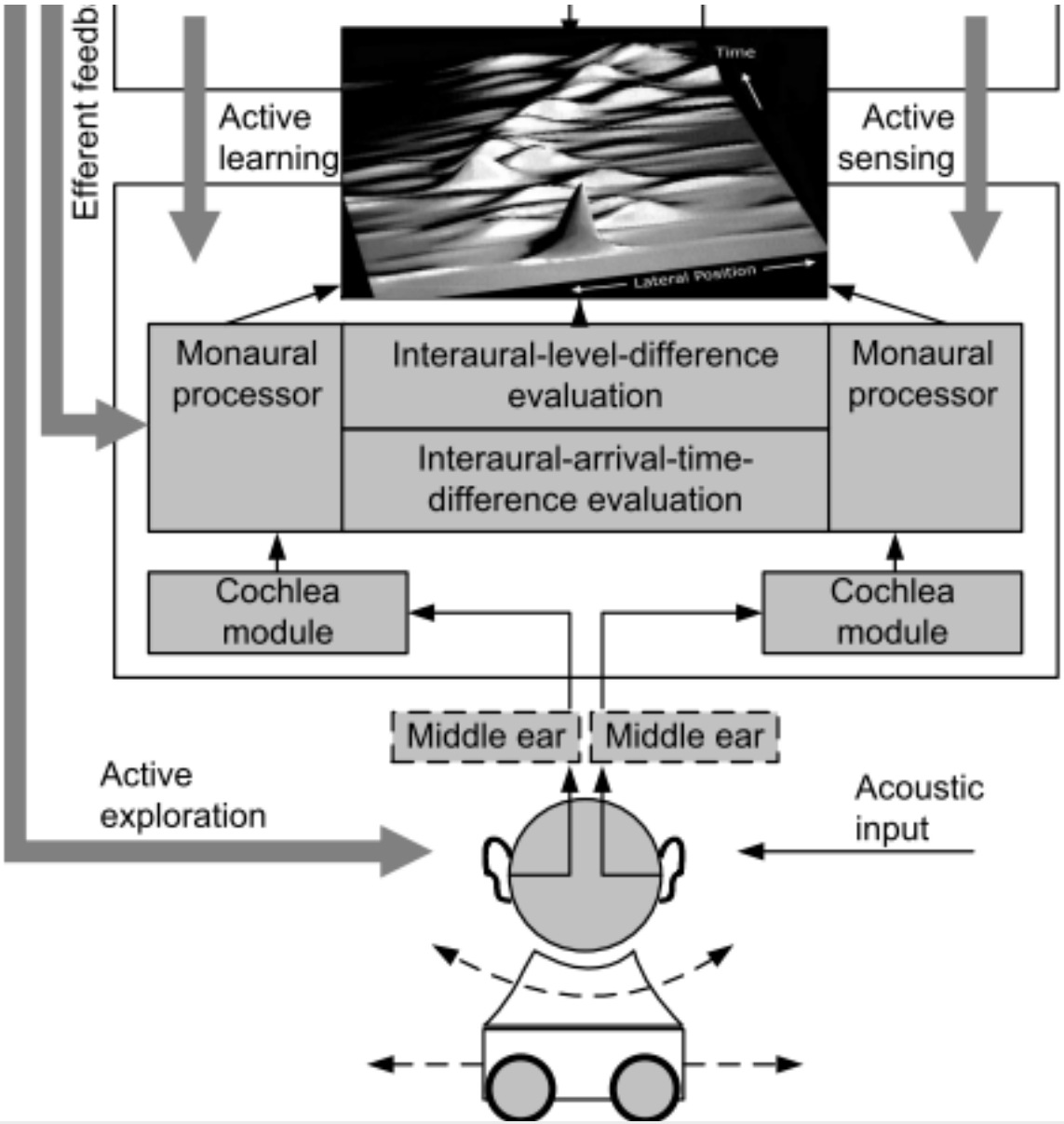
Feature Extraction

Physiologically inspired computation of

- Gammatone filterbank
- Ratemap features
- Estimation of inter-aural time and level difference
- many more...

see also http://www.twoears.eu/ (doc)

https://github.com/TWOEARS (open source, complete perceptual system)

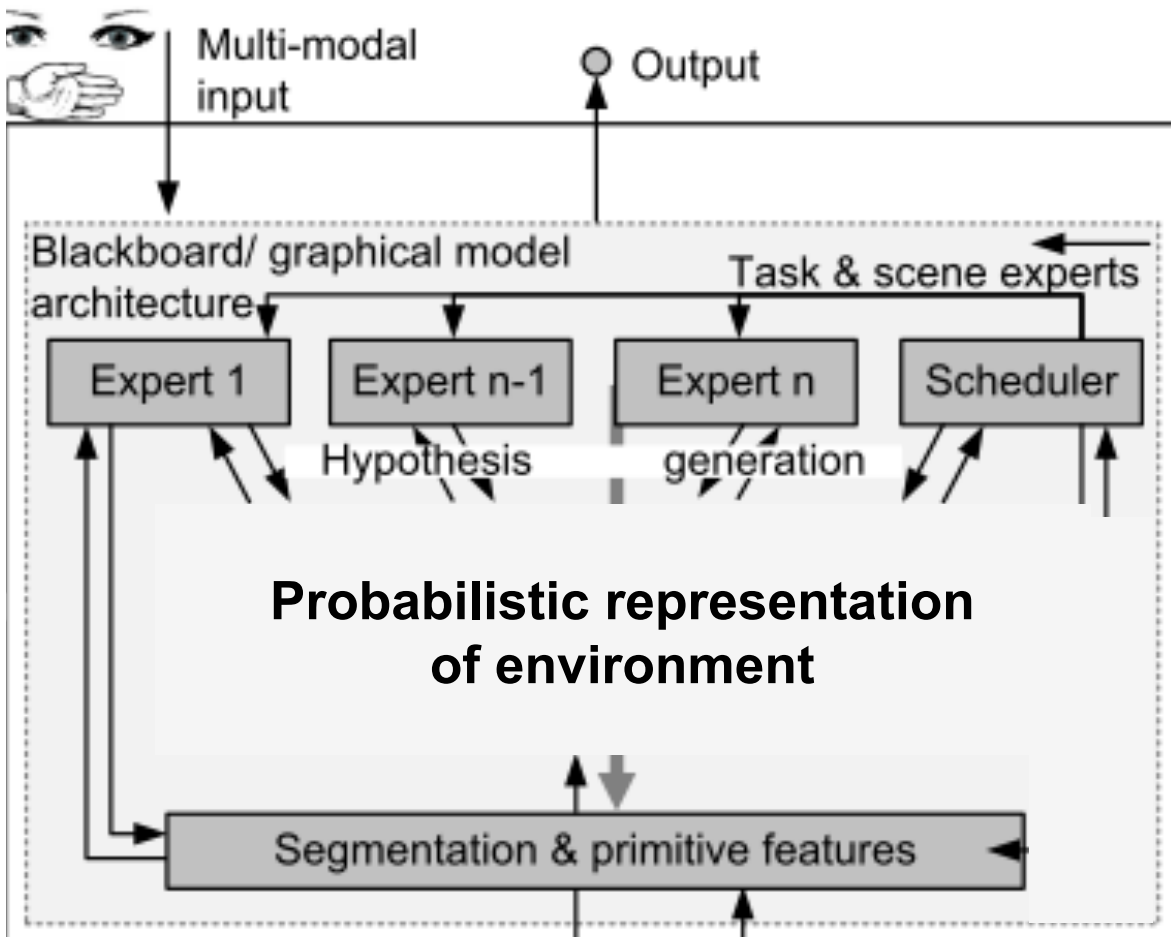


Active Perception

Cognitive Layer

Graphical-model based probabilistic blackboard architecture.

Multiple probabilistic Bayesian, DNN-based and SVM-based “experts” collaborate to form a joint probabilistic representation of the environment



[Blauert2013] J. Blauert, D. Kolossa, K. Obermayer, and K. Adiloglu: “Further challenges and the road ahead”, in J. Blauert (ed.) The technology of binaural listening, Springer, 2013.

Active Perception

Content of probabilistic blackboard

Acoustic and visual estimates of
- Source positions
- Source identities

Acoustic estimates of
- Source dominance over time and frequency (segregation hypotheses)
- Number of sources

Laser-sensor based robot position

Estimates are stored probabilistically, as histogram or density estimates.

Active Perception

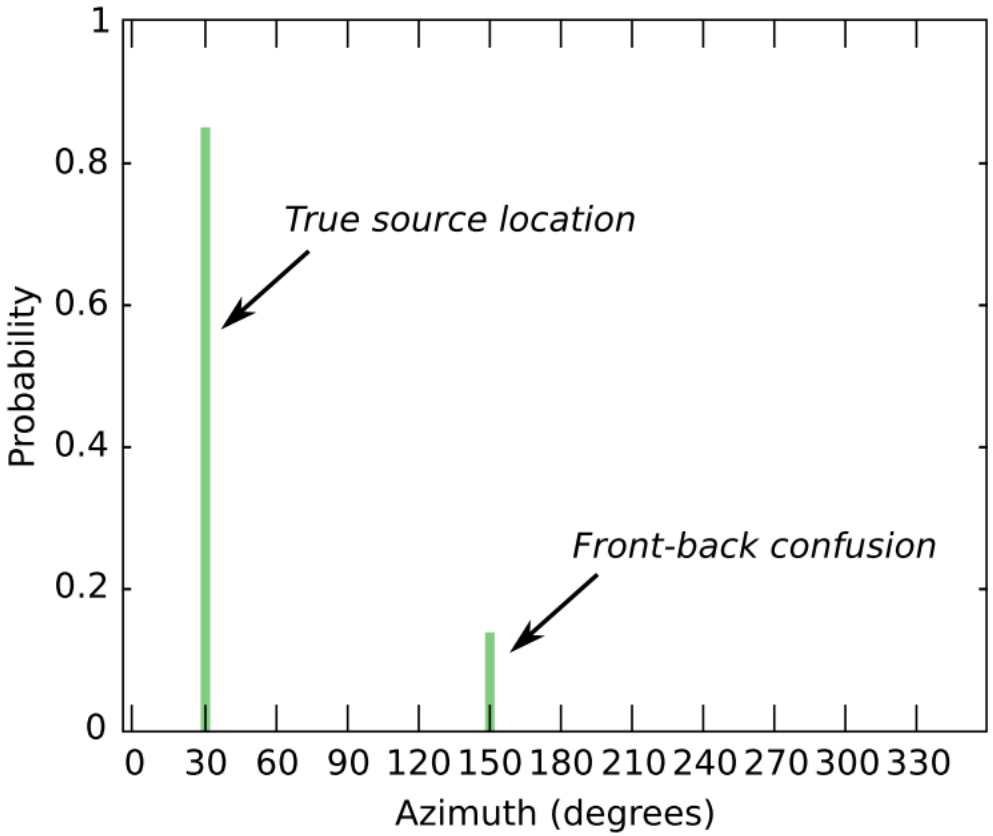
2 exemplary tasks

a) Source localization based on head rotations

b) Source localization based on robot movement

A) Active Localization through Head Rotations

Exemplary probability distribution on blackboard



Comparing 3 strategies:

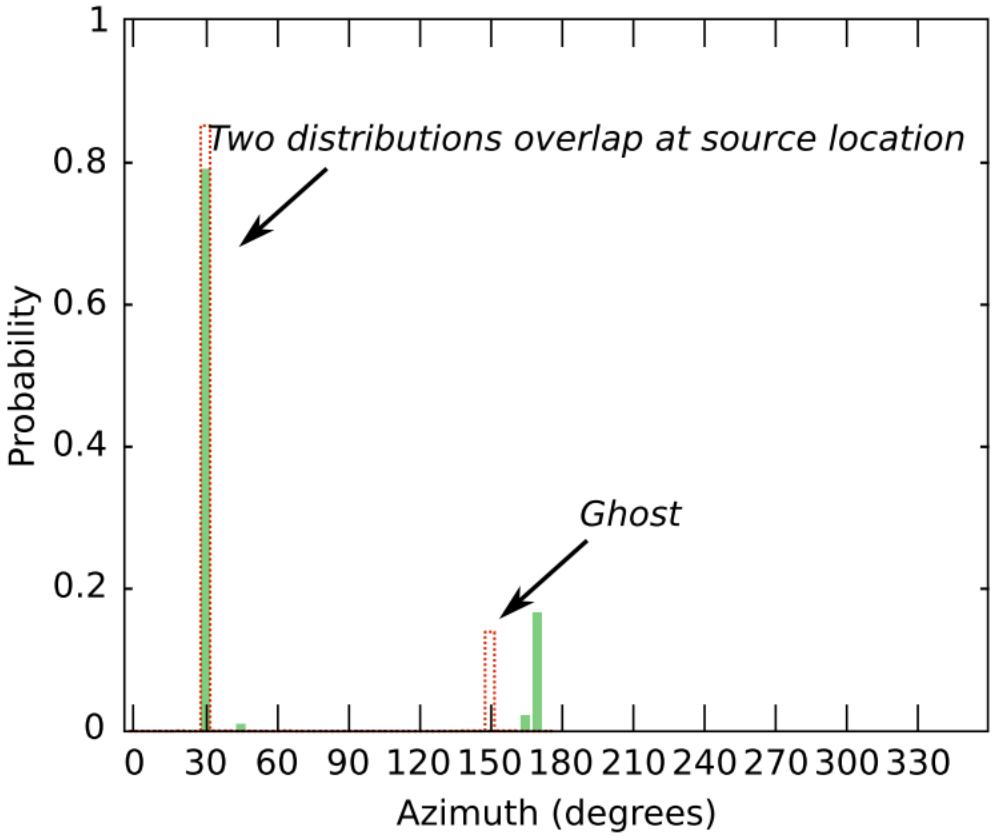
NoRot. - Passive localization by Kalman Filter, no movement

Scan - Rotate head position around 0°

PM - Smooth posterior mean, rotate head to mean of posterior distribution

A) Active Localization through Head Rotations

Exemplary probability distribution on blackboard



Comparing 3 strategies:

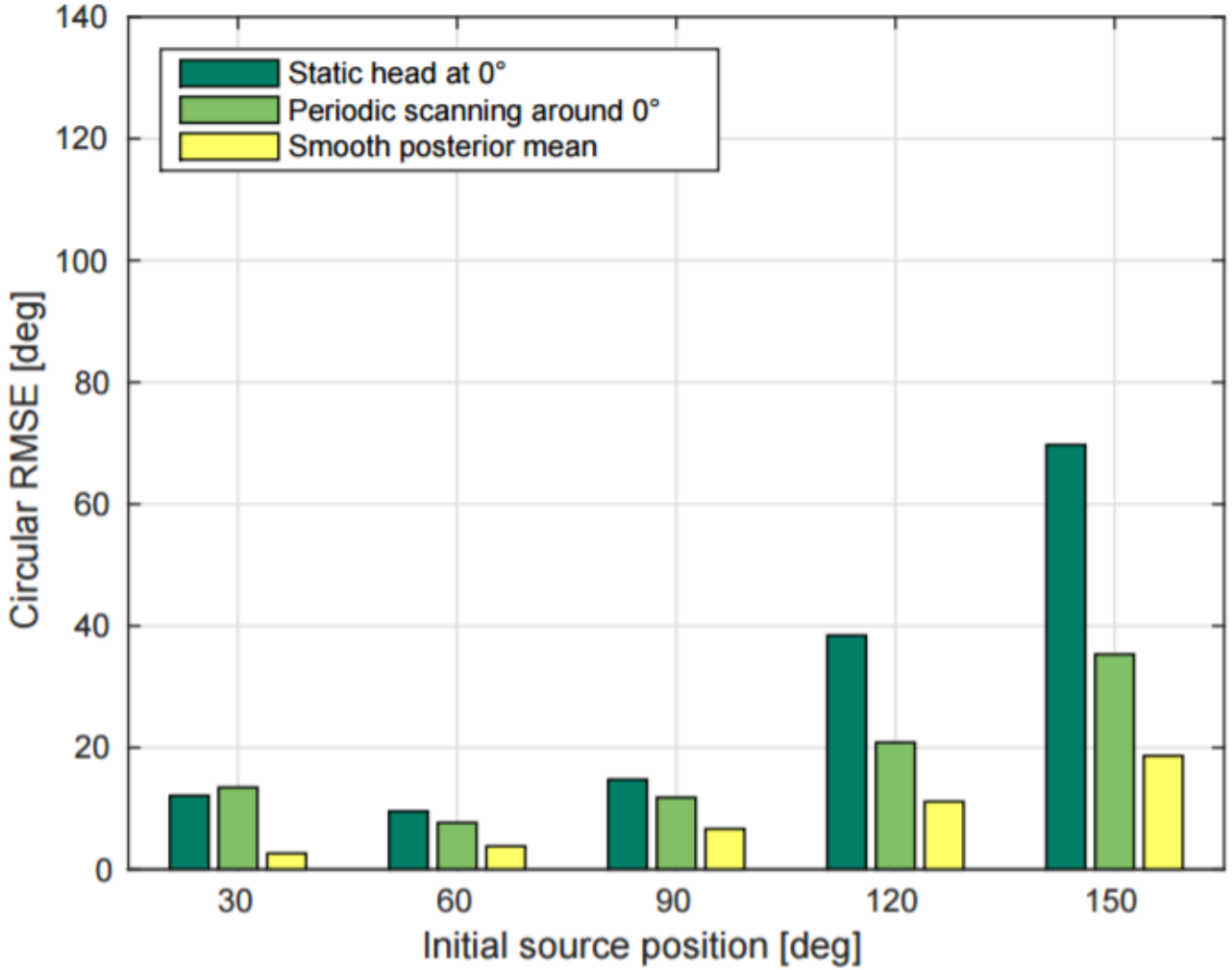
NoRot. - Passive localization by Kalman Filter, no movement

Scan - Rotate head position around 0°

PM - Smooth posterior mean, rotate head to mean of posterior distribution

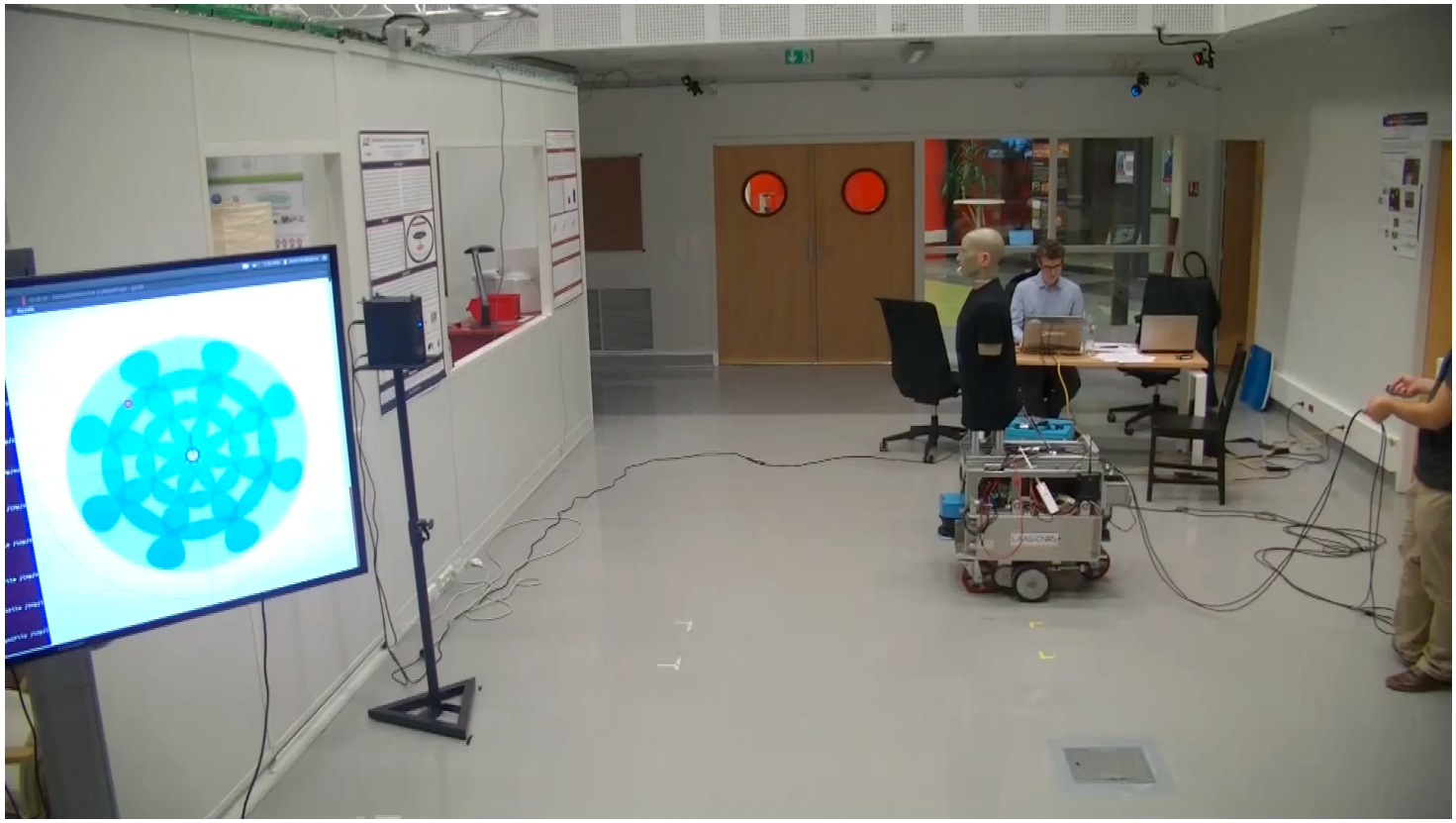
[Schymura2016] C. Schymura, J. Rios Grajales, D. Kolossa: “Active localization of sound sources with binaural models,” Proc. DAGA, Aachen, Germany, March 2016.

A) Active Localization through Head Rotations



[Schymura2015] C. Schymura, D. Kolossa, F. Winter, S. Spors: “Binaural Sound Source Localisation and Tracking using a Dynamic Spherical Head Model,” Proc. DAGA, 2015.

B) Active Localization through Movement Planning

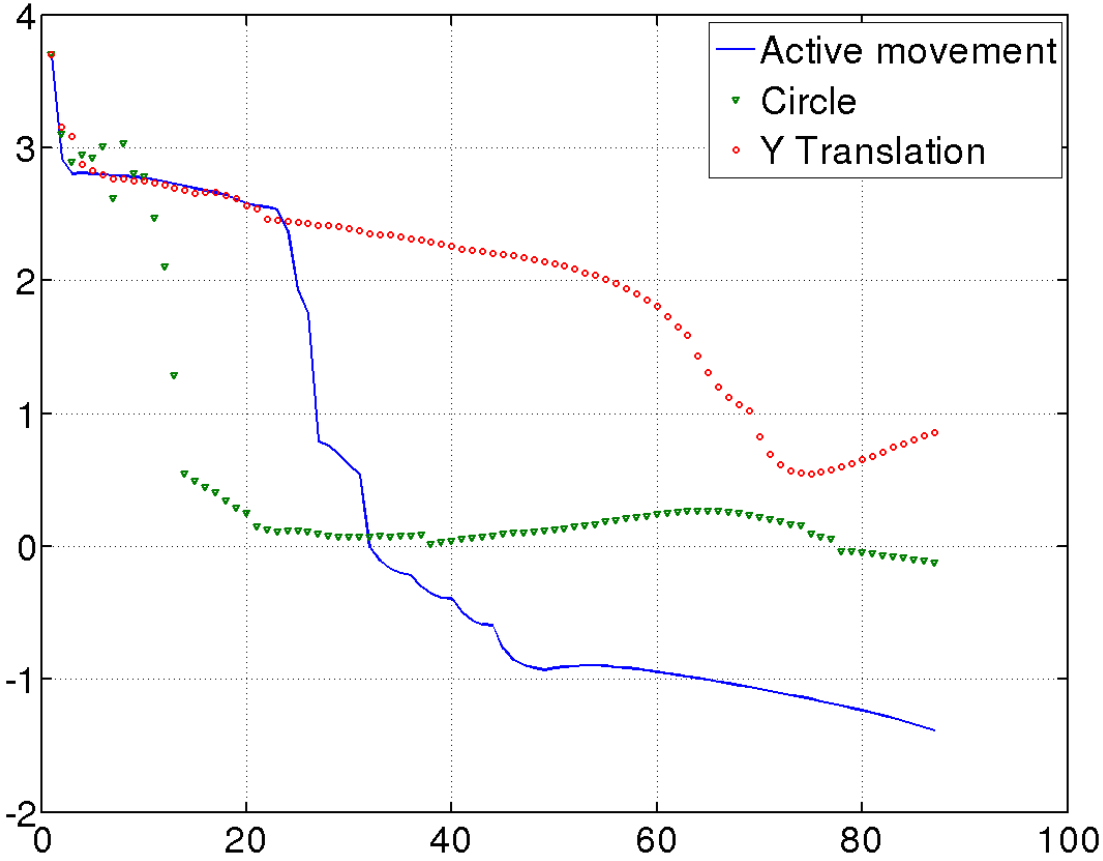


[Bustamante2016] G. Bustamante, P. Danes, T. Forgue, A. Podlubne: “Towards Information-based feedback control for binaural active localization,” Proc. ICASSP 2016.

B) Active Localization through Movement Planning

Entropy of posterior distribution over time

Comparing three movement strategies: Active movement vs. two pre-planned trajectories



[Bustamante2016] G. Bustamante, P. Danes, T. Forgue, A. Podlubne: “Towards Information-based feedback control for binaural active localization,” Proc. ICASSP 2016.

Conclusions and Future Work

Conclusions

Bayesian perception strategies

- allow systems to incrementally construct a probabilistic description of themselves in their environment.
- Bayesian integration can be used to fuse multiple sources of information, multiple features, and observations at multiple points in time, e.g. for more robust speech recognition.
- Active perception strategies can use this Bayesian representation for optimum movement planning (in the sense of minimizing residual entropy/uncertainty).

Future Work

Next questions

- How are uncertainties represented in neural networks of mammals and how does this translate to processing uncertain information in neural-network-based learning and inference?
- How can we best learn the uncertainties of observations from data?
- And can (and should) we construct joint and integrative representations of the self in the environment, making the best use of the strengths and weaknesses of statistical, kernel-based, ensemble-learning and neural methods?

...and many thanks for your attention!

References

[Abdelaziz2013] A. Hussen Abdelaziz, S. Zeiler, D. Kolossa, V. Leutnant, R. Haeb-Umbach: „GMM-based Significance Decoding“, Proc. ICASSP 2013.

[Abdelaziz2015] A. Hussen Abdelaziz, S. Watanabe, J. Hershey, E. Vincent, D. Kolossa : “On Decoding of Uncertain Data Using Deep Neural Networks,” Proc. Interspeech 2015

[Astudillo2013] Ramón Fernández Astudillo and Reinhold Orglmeister: “Computing MMSE Estimates and Residual Uncertainty Directly in the Feature Domain of ASR using STFT Domain Speech Distortion Models” IEEE Trans. Audio Speech and Language Processing, Vol. 21, No. 5, May 2013.

[Barker2007] Jon Barker , Martin Cooke: “Modelling speaker intelligibility in noise”, Speech Communication (49), 2007, pp. 402–417.

[Blauert2013] J. Blauert, D. Kolossa, K. Obermayer, and K. Adiloglu: “Further challenges and the road ahead”, in J. Blauert (ed.) The technology of binaural listening, Springer, 2013.

[Boll1979] S. Boll: “Suppression of speech in noise using spectral subtraction” Proc. IEEE AASP-27, April 1979, pp. 113-120.

[Bustamante2016] Gabriel Bustamante, Patrick Danes, Thomas Forgue, Ariel Podlubne: “Towards Information-based feedback control for binaural active localization,” Proc. ICASSP 2016.

[Cheng2007] Cheng, Huttenlocher, Shettleworth and Rieser: “Bayesian Integration of Spatial Information,” Psychological Bulletin, vol 133, no. 4, 2007.

[Cohen2003] Israel Cohen: “Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging” IEEE Trans. Speech and Audio Processing, Vol. 11, No. 5, Sept. 2003, pp. 466-475.

[Cooke2001] M.P. Cooke, P. D. Green, L. Josifovski, A. Vizinho, A.: “Robust automatic speech recognition with missing and uncertain acoustic data,” Speech Communication: https://doi.org/SCOMDH 2001, pp. 267–285.

References

[Cooke2006] M.P. Cooke: “A glimpsing model of speech perception in noise,” J. Acoust. Soc. of America, vol. 199, pp. 1562-1573.

[Deng2005] L. Deng, J. Droppo and A. Acero: „Dynamic Compensation of HMM Variances using a feature enhancement uncertainty computed from a parametric model of speech distortion“, IEEE Trans. Speech and Audio Processing, 13(3), May 2005.

[Drude2015] L. Drude, A. Chinaev, R. Haeb-Umbach: “BLSTM-Supported GEV Beamformer Front-End for the 3rd CHiME Challenge,” Proc. ASRU 2015.

[Ernst2002] M. Ernst and M. Banks: “Humans integrate visual and haptic information in a statistically optimal fashion,” Nature, vol 415, 2002.

[Ferreira2013] J. Ferreira, J. Lobo, P. Bessiere, M. Castelo-Branco, and J. Dias: “A Bayesian Framework for Active Artificial Perception,” IEEE Trans. Cybernetics, vol. 43, Apr. 2013.

[Gemmeke2010] Gemmeke, J. F., Remes, U. and Palomäki, K. J. (2010). Observation uncertainty measures for sparse imputation, Proc. Interspeech 2010, Chiba, Japan, pp. 2262–2265.

[Ion2008] Ion, V., Haeb-Umbach, R.: “A novel uncertainty decoding rule with applications to transmission error robust speech recognition,” IEEE Trans. Audio, Speech, and Language Processing vol 16, 2008, pp. 1047–1060.

[Kallasjoki2015] Heikki Kallasjoki: “Feature Enhancement and Uncertainty Estimation for Recognition of Noisy and Reverberant Speech ,” PhD Thesis, Aalto University, 2015.

[Knill2004] Knill and Pouget: “The Bayesian brain: the role of uncertainty in neural coding and computation” Trends in Neuroscience, Vol. 27, No. 12, December 2004.

[Kolossa2005] D. Kolossa, A. Klimas, R. Orglmeister: „Separation and Recognition of Noisy, Convolutive Speech Mixtures using Time-Frequency Masking and Missing Data Techniques“, Proc. Waspaa 2005.

References

[Kolossa2010] D. Kolossa, R. Fernandez Astudillo, E. Hoffmann and R. Orglmeister: „ Independent Component Analysis and Time-Frequency Masking for Speech Recognition in Multitalker Conditions“, EURASIP Journal on Audio, Speech, and Music Processing. vol. 2010, Article ID 651420, 13 pages, 2010.

[Kolossa2013] D. Kolossa, S. Zeiler, R. Saeidi, R.F. Astudillo: “Noise-Adaptive LDA: A New Approach for Speech Recognition Under Observation Uncertainty, IEEE Signal Processing Letters, vol. 20, no. 11, pp. 1018-1021, 2013.

[LeRoux2012] Jonathan Le Roux, John R. Hershey: “Indirect model-based speech enhancement,” Proc. ICASSP 2012, pp. 4045–4048.

[LeRoux2013] Jonathan Le Roux, Shinji Watanabe, John R. Hershey: “Ensemble learning for speech enhancement”, Proc. WASPAA 2013.

[Liao2005] H. Liao, M.F. Gales: “Joint Uncertainty Decoding for Noise Robust Speech Recognition,” Proc. Interspeech, 2005.

[Liutkus2011] A. Liutkus, R. Badeau, G. Richard: “Gaussian processes for underdetermined source separation” IEEE Transactions on Signal Processing, 59 (7), 2011, pp. 3155-3167.

[Ma2009] Wei Ji Ma, Xiang Zhou, Lars A. Ross, John J. Foxe, Lucas C. Parra: “Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space” PLoS ONE 4(3). 2009 doi:10.1371/journal.pone.0004638

[Martin2003] Rainer Martin: “Statistical methods for the enhancement of noisy speech,” Proc. IWAENC 2003.

[Martin2001] Rainer Martin, “Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics,” IEEE Trans. Speech and Audio Processing, vol. 9, pp. 504–512, July 2001

[Nesta2013] F. Nesta, M. Matassoni, and R. Astudillo, “A flexible spatial blind source extraction framework for robust speech recognition in noisy environments,” in Proc. CHiME, 2013, pp. 33–40.

[Raj2005] B. Raj and R. Stern:” Missing-Feature Approaches in Speech Recognition,” IEEE Signal Processing Magazine, Sept. 2005, pp. 101-116.

References

[Srinivasan2007] S. Srinivasan and D. Wang: “Transforming binary uncertainties for robust speech recognition, IEEE Transactions on Audio, Speech, and Language Processing” 15(7), 2007, pp. 2130–2140.

[Schymura2014] Ch. Schymura C. Schymura, T. Walther, D. Kolossa, N. Ma and G. Brown: “Binaural Sound Source Localisation using a Bayesian-network-based Blackboard System and Hypothesis-driven Feedback,” Proc. Forum Acusticum, Krakow, September 2014.

[Schymura2015] C. Schymura, D. Kolossa, F. Winter, S. Spors: “Binaural Sound Source Localisation and Tracking using a Dynamic Spherical Head Model,” Proc. DAGA, 2015.

[Schymura2016] C. Schymura, J. Rios Grajales, D. Kolossa: “Active localization of sound sources with binaural models,” Proc. DAGA, Nürnberg, March 2016

[Tachioka2013] Y. Tachioka, S. Watanabe, J. L. Roux, and J. R. Hershey: “Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark,” in The 2nd International Workshop on Machine Listening in Multisource Environments (CHiME), Vancouver, Canada, 2013.

[Tran2014] Dung Tran, Emmanuel Vincent, Denis Jouvet: “Fusion of Multiple Uncertainty Estimators and Propagators for Noise Robust ASR”. Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy, May 2014.

[Taghia2016] J. Taghia, D. Kolossa, R. Martin: „ALE for Robots! A single-channel approach to robot self noise reduction“, IWAENC 2016.

[Vincent2013] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni,“The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines,” in Proc. ICASSP, 2013, pp. 126–130.

[Vincent] E. Vincent, T. Virtanen, S. Gannot: “Audio Source Separation and Speech Enhancement” to appear, Wiley.

Projection 2: Linear Discriminant Analysis

Definitions:

Class covariance of class *c*

$$\boldsymbol{\Sigma}_c = \frac{1}{N_c} \sum_{\mathbf{x}_i \in c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top, \quad \boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\mathbf{x}_i \in c} \mathbf{x}_i$$

Between-class covariance

$$\boldsymbol{\Sigma}_b = \sum_{\forall c} \frac{N_c}{N} (\boldsymbol{\mu}_c - \bar{\mathbf{x}})(\boldsymbol{\mu}_c - \bar{\mathbf{x}})^\top, \qquad \bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

Within-class covariance

$$\boldsymbol{\Sigma}_w = \sum_{\forall c} \frac{N_c}{N} \boldsymbol{\Sigma}_c,$$

Projection 3: *Noise-Adaptive* Linear Discriminant Analysis

Basic idea: Since estimate of feature uncertainty (in the form of an estimated error covariance matrix) is available, modify class covariance matrices accordingly:

$$
\begin{aligned}
\boldsymbol{\Sigma}_{wn,c} &= \mathrm{E}\left((\hat{\mathbf{x}}_c - \boldsymbol{\mu}_c)(\hat{\mathbf{x}}_c^\top - \boldsymbol{\mu}_c^\top)\right) \\
&= \mathrm{E}\left(\hat{\mathbf{x}}_c\,\hat{\mathbf{x}}_c^\top\right) - \mathrm{E}\left(\boldsymbol{\mu}_c\,\boldsymbol{\mu}_c^\top\right) \\
&= \mathrm{E}\left((\mathbf{x}_c + \mathbf{n})(\mathbf{x}_c + \mathbf{n})^\top\right) - \mathrm{E}\left(\boldsymbol{\mu}_c\,\boldsymbol{\mu}_c^\top\right) \\
&= \underbrace{\mathrm{E}\left(\mathbf{x}_c\,\mathbf{x}_c^\top\right) - \mathrm{E}\left(\boldsymbol{\mu}_c\,\boldsymbol{\mu}_c^\top\right)}_{\boldsymbol{\Sigma}_{w,c}} + \mathrm{E}\left(\mathbf{n}\,\mathbf{n}^\top\right) \\
&= \boldsymbol{\Sigma}_{w,c} + \mathrm{E}\left(\mathbf{n}\,\mathbf{n}^\top\right) \\
&\approx \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_n(\tau)
\end{aligned}
$$

D. Kolossa, S. Zeiler, R. Saeidi, R.F. Astudillo: “Noise-Adaptive LDA: A New Approach for Speech Recognition Under Observation Uncertainty, IEEE Signal Processing Letters, vol. 20, no. 11, pp. 1018-1021, 2013.

The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind." James Clerk Maxwell (1850)

from E.T. Jaynes, Probability Theory The Logic of Science, Cambridge University Press 2003.

Example for Single Channel Case

Uncertainty Estimation

Use residual error of MMSE-Estimator

$$\mathrm{Var}\{X_{fn} \mid Y\} = \frac{\lambda_X \lambda_D}{\lambda_X + \lambda_D}$$

or Bernoulli model of uncertainty

$$\mathrm{Var}\{X_{fn} \mid Y\} = \hat{p}_{fn}(1 - \hat{p}_{fn})Y_{fn}^2 \text{ with } \hat{p}_{fn} = \frac{\sqrt{\lambda_X}}{\sqrt{\lambda_X} + \sqrt{\lambda_D}}$$

[1] Ramón Fernández Astudillo and Reinhold Orglmeister: “Computing MMSE Estimates and Residual Uncertainty Directly in the Feature Domain of ASR using STFT Domain Speech Distortion Models” IEEE Trans. Audio Speech and Language Processing, Vol. 21, No. 5, May 2013.

[2] F. Nesta, M. Matassoni, and R. Astudillo, “A flexible spatial blind source extraction framework for robust speech recognition in noisy environments,” in Proc. CHiME, 2013, pp. 33–40.