CARL VON OSSIETZKY universität OLDENBURG

Parallel Particle Filters for Tracking of Voice Glimpses

Joanna Luberadzka, Hendrik Kayser, Volker Hohmann
Digital Hearing Devices,
Department of Medical Physics and Acoustics,
University of Oldenburg

CARL VON OSSIETZKY universität OLDENBURG

Presentation:

I. INTRODUCTION

I. 1. MOTIVATION
I. 2. PSYCHOACOUSTIC STUDY
I. 3. MODELING FRAMEWORK
I. 4. SIMULATION APPROACH

II. METHODS

II. 1. TRAJECTORY GENERATION
II. 2. OBSERVATION GENERATION
II. 3. PARTICLE FILTERING
II. 3. a. Initialization
II. 3. b. State prediction
II. 3. c. Glimpse association
II. 3. d. Weight update
II. 3. e. Current state estimation
II. 3. f. Resampling

III. SIMULATION

III.1. EXPERIMENT DESIGN
III.2. RESULTS
III.3. CONCLUSIONS
III.4. OUTLOOK

1. INTRODUCTION

I. INTRODUCTION

I. 1. MOTIVATION

I. 1. MOTIVATION

CARL VON OSSIETZKY universität OLDENBURG

I. 1. MOTIVATION

- cocktail party

- challenging acoustic conditions, many competing voices

- but still we are able to follow a desired acoustic object

I. 1. MOTIVATION

- cocktail party

- challenging acoustic conditions, many competing voices

- but still we are able to follow a desired acoustic object

- understand this phenomenon

- we know: pitch, timbre, location help to form streams

- but what if they change in time? can we still follow?

I. INTRODUCTION

I. 2. PSYCHOACOUSTIC STUDY

I. 2. PSYCHOACOUSTIC STUDY

CARL VON OSSIETZKY universität OLDENBURG

Attentive Tracking of Sound Sources

Kevin J.P. Woods[1,2,*] and Josh H. McDermott[1,2]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA 02138, USA
*Correspondence: kwoods@mit.edu
http://dx.doi.org/10.1016/j.cub.2015.07.043

SUMMARY

Auditory scenes often contain concurrent sound sources, but listeners are typically interested in just one of these and must somehow select it for further processing. One challenge is that real-world sounds such as speech vary over time and as a consequence

[1, 13–16], but less is known about the processes underlying attentional selection and their interaction with sound segregation [17–19].

Both segregation and selection could be aided by features of a target source that distinguish it from other sources, such as a unique pitch or location [20–23]. Studies of stream segregation have largely focused on cases such as this [1, 13–16, 24–30],

I. 2. PSYCHOACOUSTIC STUDY

CARL VON OSSIETZKY universität OLDENBURG

Attentive Tracking of Sound Sources

Kevin J.P. Woods[1,2,*] and Josh H. McDermott[1,2]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA 02138, USA
*Correspondence: kwoods@mit.edu
http://dx.doi.org/10.1016/j.cub.2015.07.043

SUMMARY

Auditory scenes often contain concurrent sound sources, but listeners are typically interested in just one of these and must somehow select it for further processing. One challenge is that real-world sounds such as speech vary over time and as a consequence

[1, 13–16], but less is known about the processes underlying attentional selection and their interaction with sound segregation [17–19].

Both segregation and selection could be aided by features of a target source that distinguish it from other sources, such as a unique pitch or location [20–23]. Studies of stream segregation have largely focused on cases such as this [1, 13–16, 24–30],

- **reduced** cocktail party to two simultaneously active '**singing**' voices (only vowels)

- voices synthesized: **no constant differences** for example in timbre

I. 2. PSYCHOACOUSTIC STUDY

CARL VON OSSIETZKY universität OLDENBURG

Attentive Tracking of Sound Sources

Kevin J.P. Woods[1,2,*] and Josh H. McDermott[1,2]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA 02138, USA
*Correspondence: kwoods@mit.edu
http://dx.doi.org/10.1016/j.cub.2015.07.043

SUMMARY

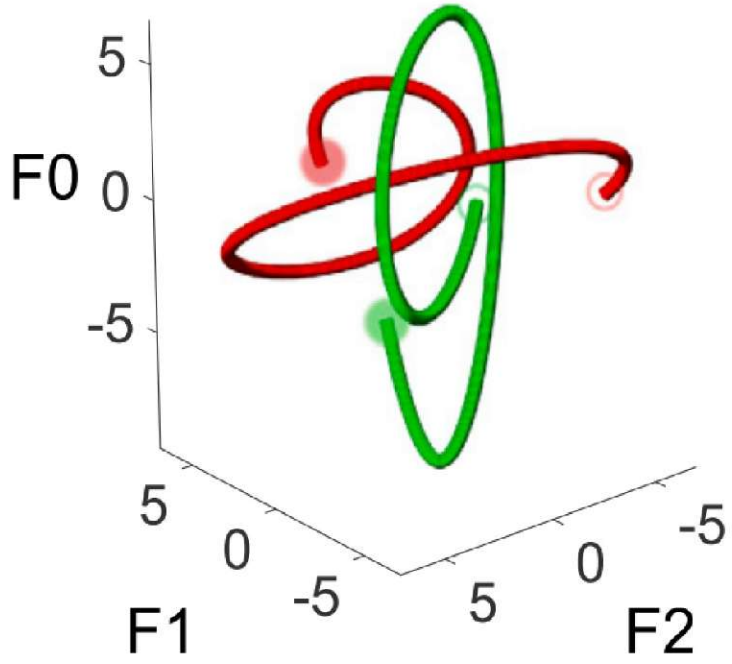
Auditory scenes often contain concurrent sound sources, but listeners are typically interested in just one of these and must somehow select it for further processing. One challenge is that real-world sounds such as speech vary over time and as a consequence

[1, 13–16], but less is known about the processes underlying attentional selection and their interaction with sound segregation [17–19].

Both segregation and selection could be aided by features of a target source that distinguish it from other sources, such as a unique pitch or location [20–23]. Studies of stream segregation have largely focused on cases such as this [1, 13–16, 24–30],

- only thing varied: how parameters(formants) **change in time**

- parameter trajectories might **cross**



I. 2. PSYCHOACOUSTIC STUDY

CARL VON OSSIETZKY universität OLDENBURG

Attentive Tracking of Sound Sources

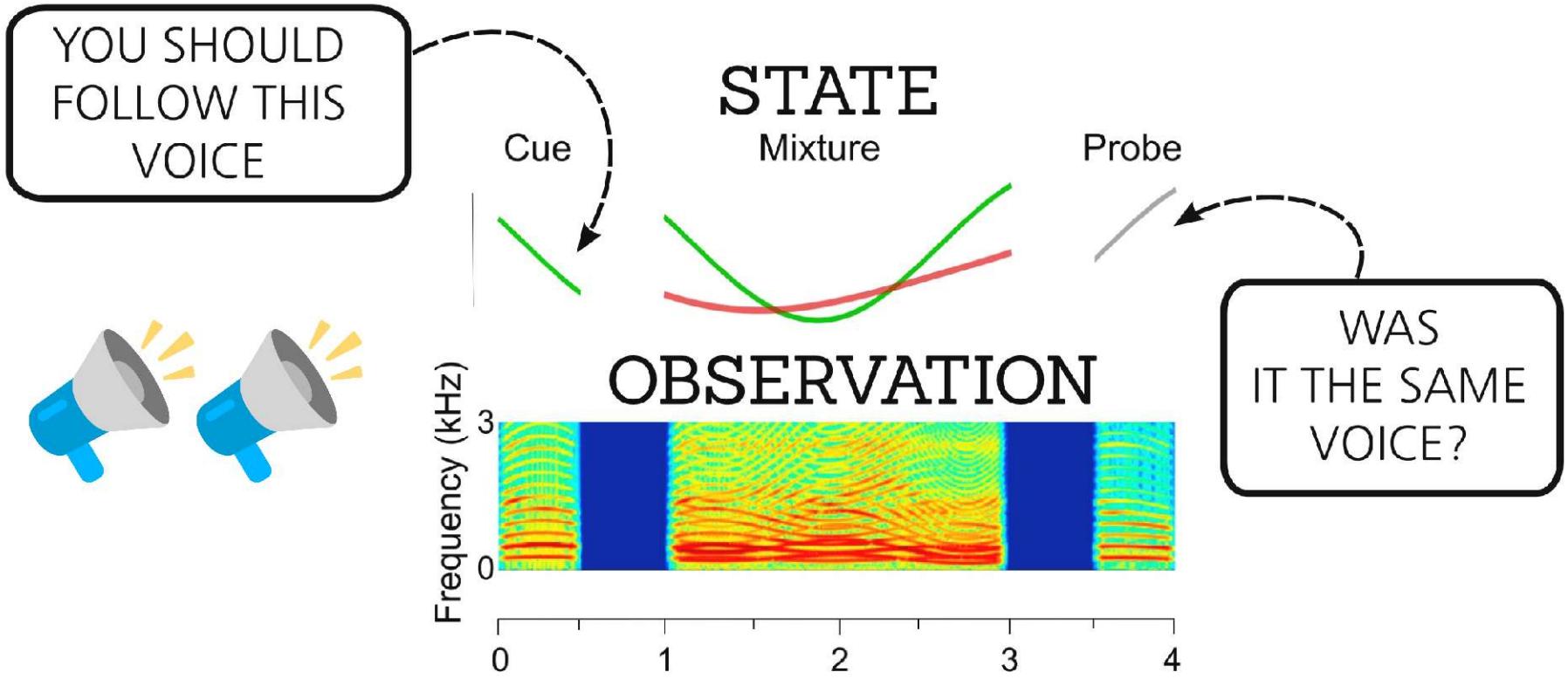
Kevin J.P. Woods[1,2,*] and Josh H. McDermott[1,2]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA 02138, USA
*Correspondence: kwoods@mit.edu
http://dx.doi.org/10.1016/j.cub.2015.07.043

SUMMARY

Auditory scenes often contain concurrent sound sources, but listeners are typically interested in just one of these and must somehow select it for further processing. One challenge is that real-world sounds such as speech vary over time and as a consequence

[1, 13–16], but less is known about the processes underlying attentional selection and their interaction with sound segregation [17–19].

Both segregation and selection could be aided by features of a target source that distinguish it from other sources, such as a unique pitch or location [20–23]. Studies of stream segregation have largely focused on cases such as this [1, 13–16, 24–30],



I. 2. PSYCHOACOUSTIC STUDY

CARL VON OSSIETZKY universität OLDENBURG

Attentive Tracking of Sound Sources

Kevin J.P. Woods[1,2,*] and Josh H. McDermott[1,2]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA 02138, USA
*Correspondence: kwoods@mit.edu
http://dx.doi.org/10.1016/j.cub.2015.07.043

SUMMARY

Auditory scenes often contain concurrent sound sources, but listeners are typically interested in just one of these and must somehow select it for further processing. One challenge is that real-world sounds such as speech vary over time and as a consequence

[1, 13–16], but less is known about the processes underlying attentional selection and their interaction with sound segregation [17–19].

Both segregation and selection could be aided by features of a target source that distinguish it from other sources, such as a unique pitch or location [20–23]. Studies of stream segregation have largely focused on cases such as this [1, 13–16, 24–30],

- **Question**: is it possible to follow one of two voices in the absence of constant dissimilarities between them?

- different conditions (including signals with speech-like **pauses**)

I. 2. PSYCHOACOUSTIC STUDY

CARL VON OSSIETZKY universität OLDENBURG

Attentive Tracking of Sound Sources

Kevin J.P. Woods[1,2,*] and Josh H. McDermott[1,2]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA 02138, USA
*Correspondence: kwoods@mit.edu
http://dx.doi.org/10.1016/j.cub.2015.07.043

SUMMARY

Auditory scenes often contain concurrent sound sources, but listeners are typically interested in just one of these and must somehow select it for further processing. One challenge is that real-world sounds such as speech vary over time and as a consequence

[1, 13–16], but less is known about the processes underlying attentional selection and their interaction with sound segregation [17–19].

Both segregation and selection could be aided by features of a target source that distinguish it from other sources, such as a unique pitch or location [20–23]. Studies of stream segregation have largely focused on cases such as this [1, 13–16, 24–30],

- **Question**: is it possible to follow one of two voices in the absence of constant dissimilarities between them?

- different conditions (including signals with speech-like **pauses**)

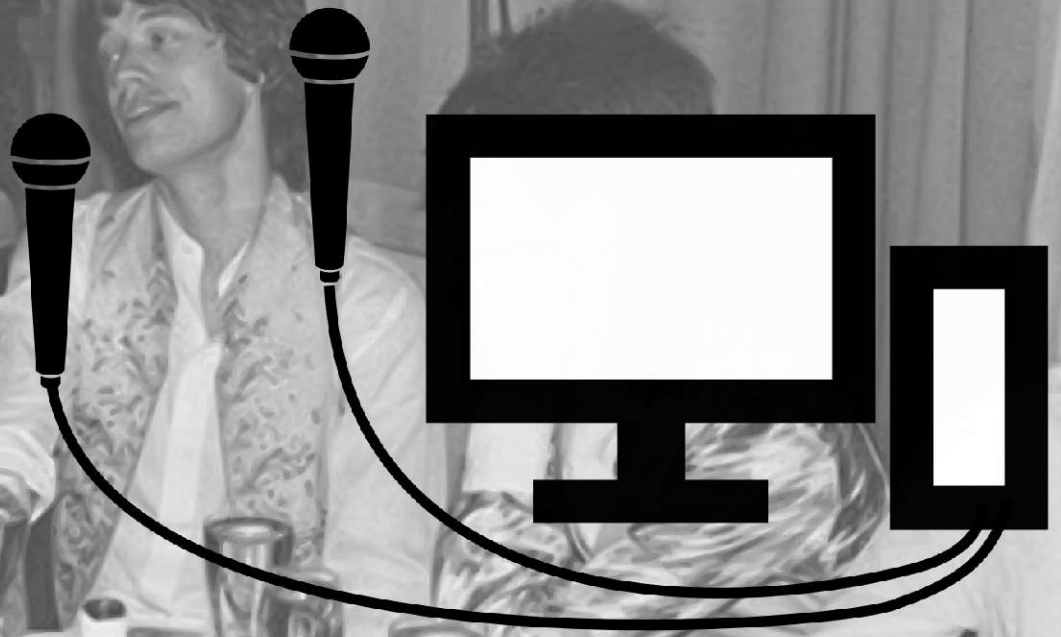
-> Result : **YES**, it is possible to discriminate if the probe belongs to the cued voice

-> **ATTENTIVE TRACKING** - moving locus of attention

I. 3. PSYCHOACOUSTIC STUDY

CARL VON OSSIETZKY universität OLDENBURG

Goal: computational model of attentive tracking task



I. INTRODUCTION

I. 3. MODELING FRAMEWORK

I. 3. MODELING FRAMEWORK

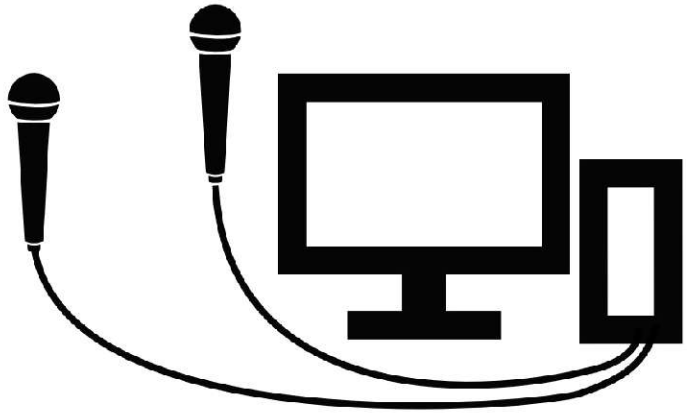
CARL VON OSSIETZKY universität OLDENBURG

ASA

Phenomenon that we want to model - **Auditory Scene Analysis**

CASA:

Computational procedures simulating this phenomenon - **Computational Auditory Scene Analysis**



I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

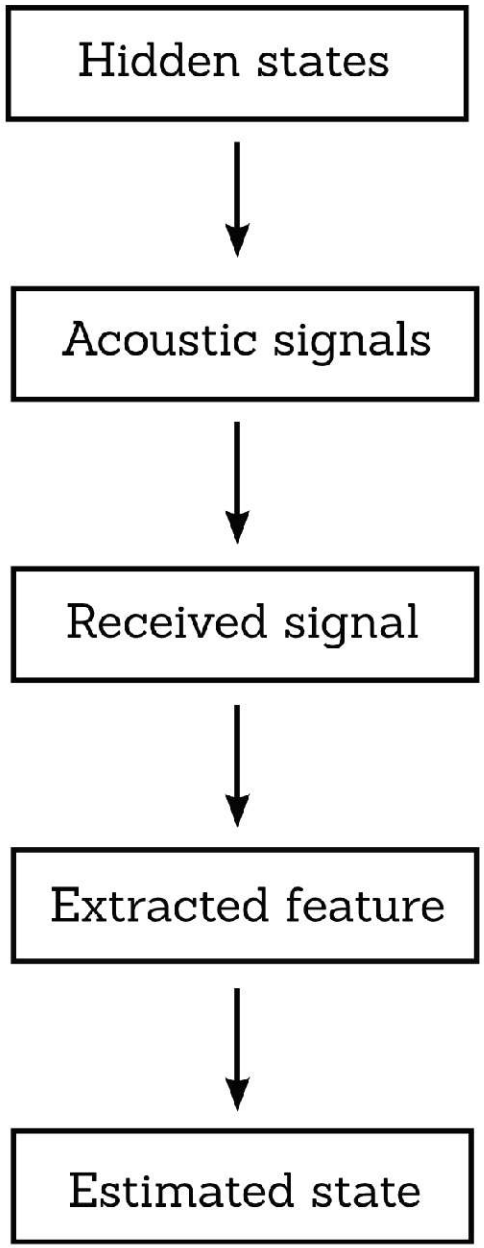
ASA

- What do I want to model?

- Block diagram ->

- General information processing / signal path

- reflected both in CASA and ASA

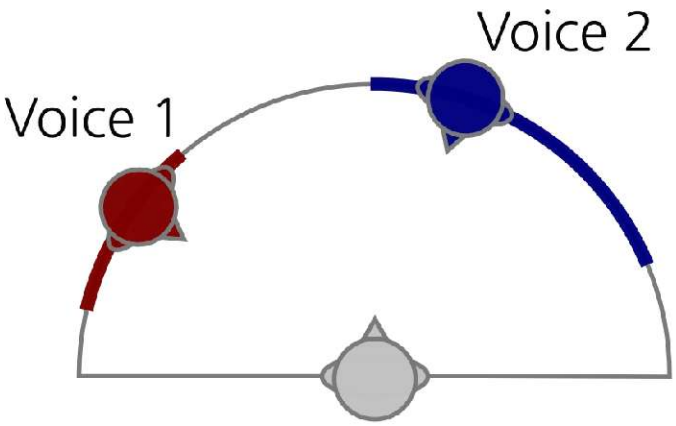


CASA:

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA



acoustic environment
two moving sources,
one receiver

REAL STATE OF A SYSTEM:
description of what is realy happening

Hidden states

↓

Acoustic signals

↓

Received signal

↓

Extracted feature

↓

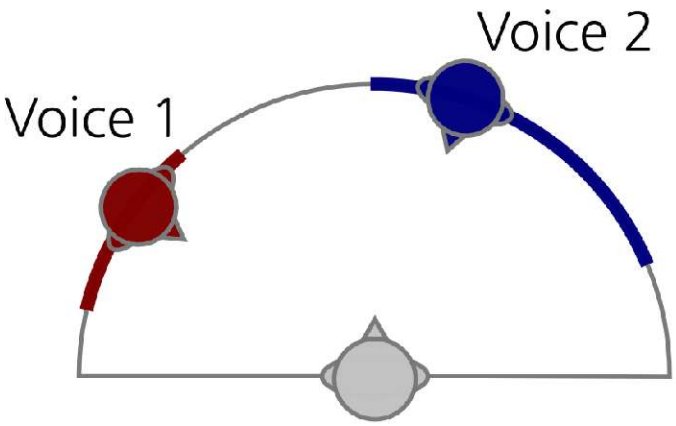
Estimated state

CASA:

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA



acoustic environment
two moving sources,
one receiver

STATE OF A SYSTEM:
description of what is happening in reality

Hidden states

↓

Acoustic signals

↓

Received signal

↓

Extracted feature

↓

Estimated state

CASA:

multidimensional state vectors with high level parameters

$$\vec{s}_{1,n} = \begin{pmatrix} F0_{1,n} \\ F1_{1,n} \\ F2_{1,n} \\ \alpha_{1,n} \end{pmatrix}$$ Voice 1

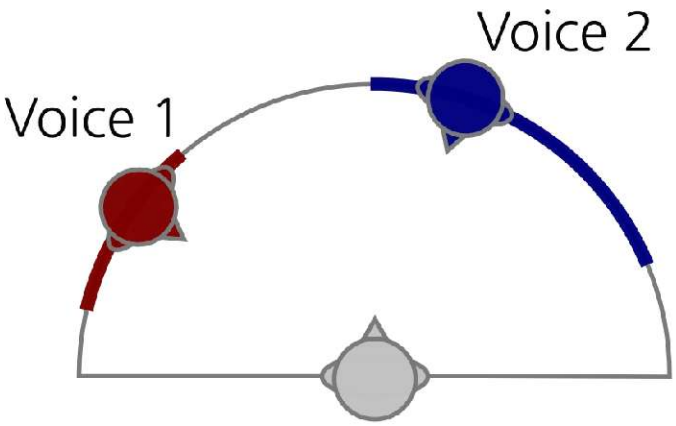
$$\vec{s}_{2,n} = \begin{pmatrix} F0_{2,n} \\ F1_{2,n} \\ F2_{2,n} \\ \alpha_{2,n} \end{pmatrix}$$ Voice 2

GROUND TRUTH STATE

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA



acoustic environment
two moving sources,
one receiver

STATE OF A SYSTEM:
description of what is happening in reality

Hidden states

↓

Acoustic signals

↓

Received signal

↓

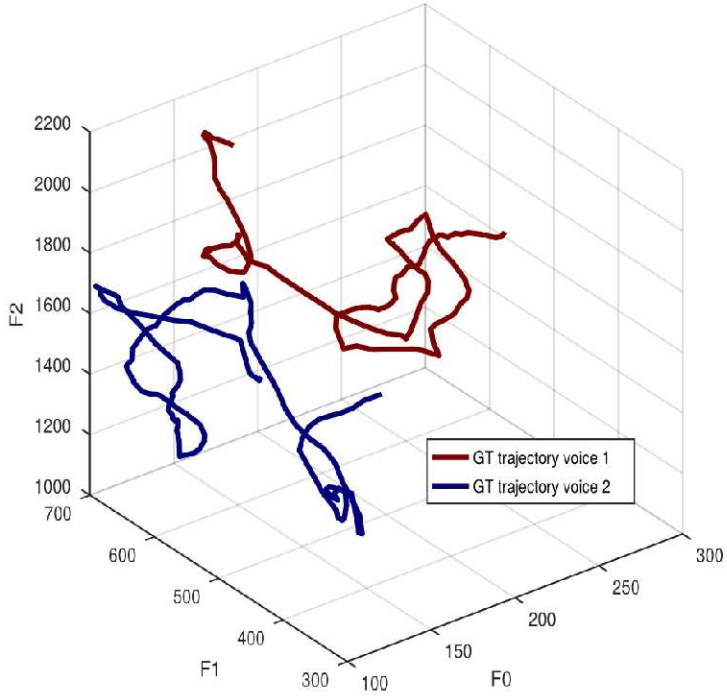
Extracted feature

↓

Estimated state

CASA:

state in time :
state trajectories

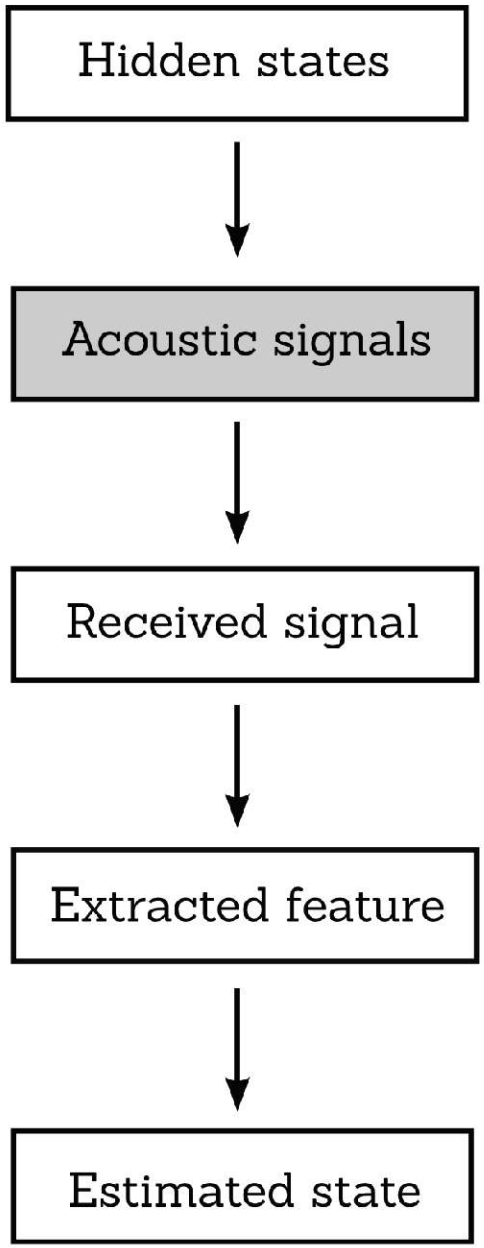


GROUND TRUTH STATE TRAJECTORY

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

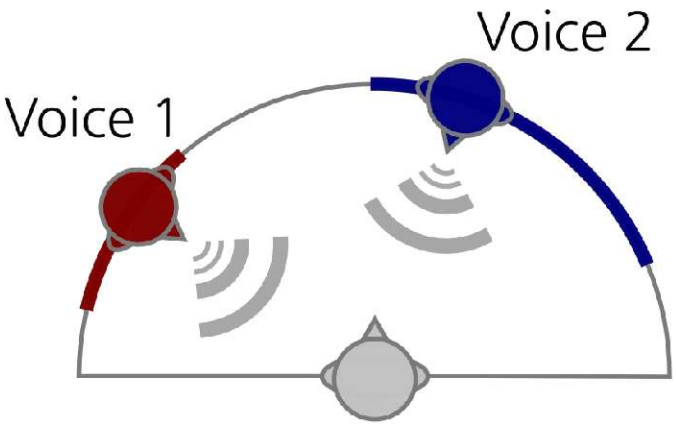


CASA:

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA



Sources in the acoustic environment generate sound pressure waves

Hidden states

↓

Acoustic signals

↓

Received signal

↓

Extracted feature

↓

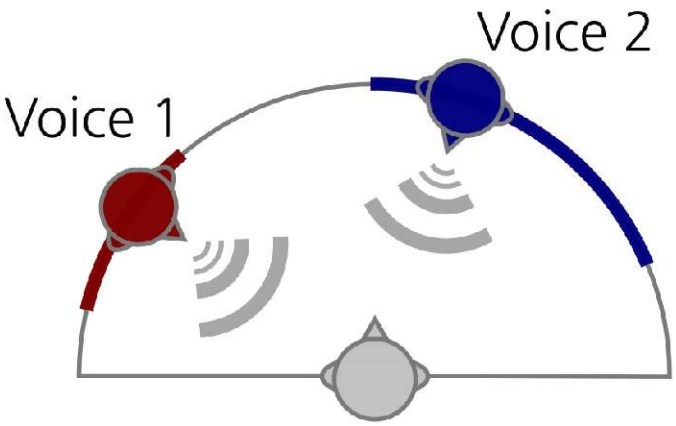
Estimated state

CASA:

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA



Sources in the acoustic environment generate sound pressure waves

Hidden states

↓

Acoustic signals

↓

Received signal

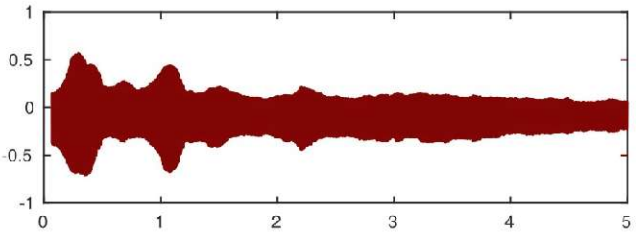
↓

Extracted feature

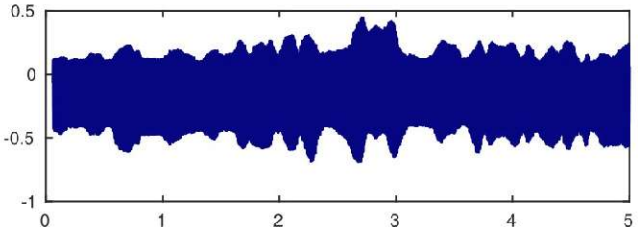
↓

Estimated state

CASA:



Voice 1



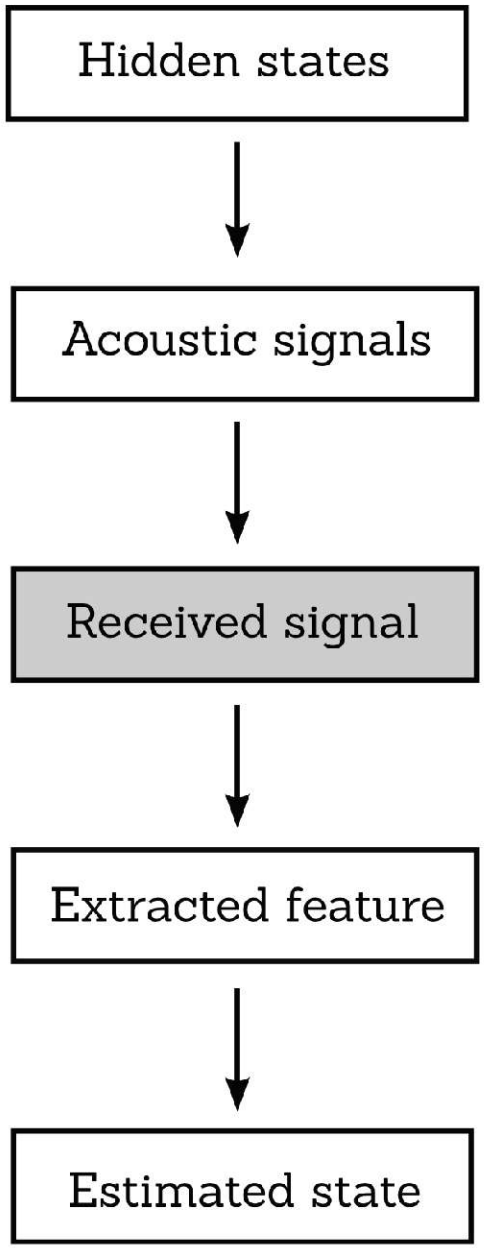
Voice 2

Waveforms of the generated sounds in a digital form

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

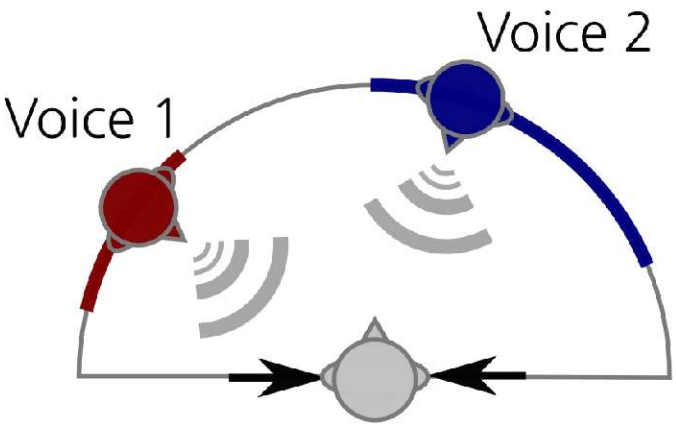


CASA:

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA



Listener - the interfering **mixture** of two sound waves makes the eardrums vibrate

CASA:

Hidden states

↓

Acoustic signals

↓

Received signal

↓

Extracted feature

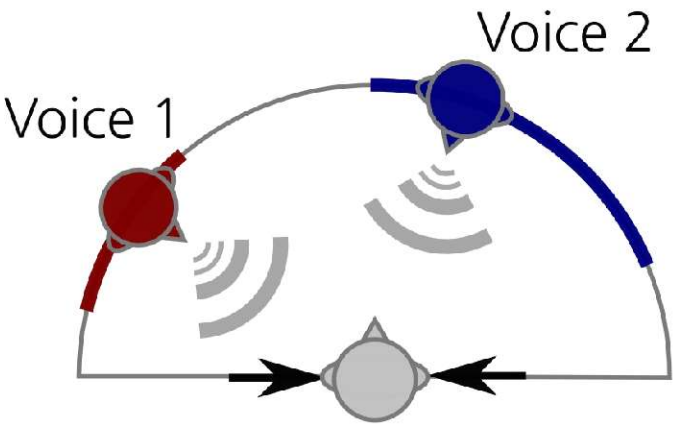
↓

Estimated state

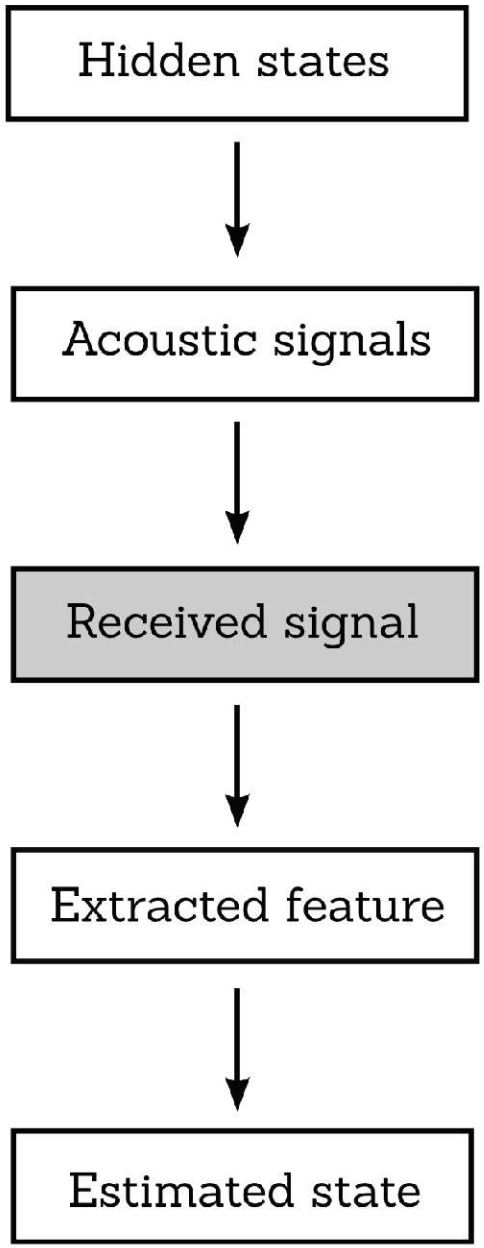
I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

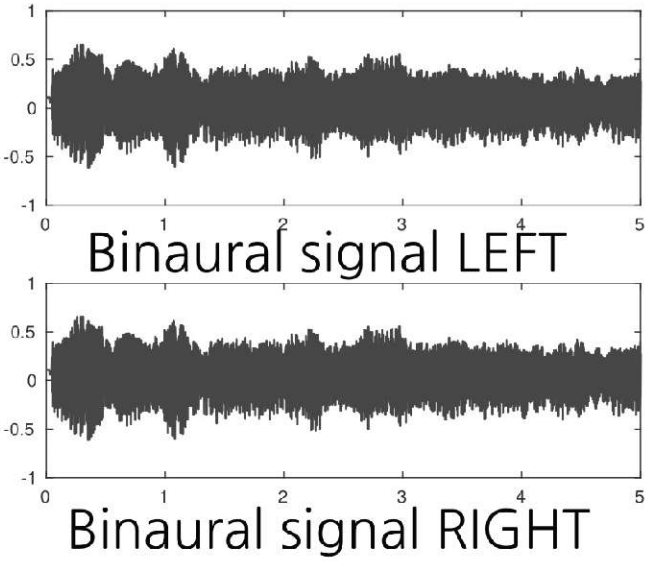
ASA



Listener - the interfering **mixture** of two sound waves makes the eardrums vibrate



CASA:



Binaural receiver -> **sum** of both signals **convolved with HRIRs** of left and right ear

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

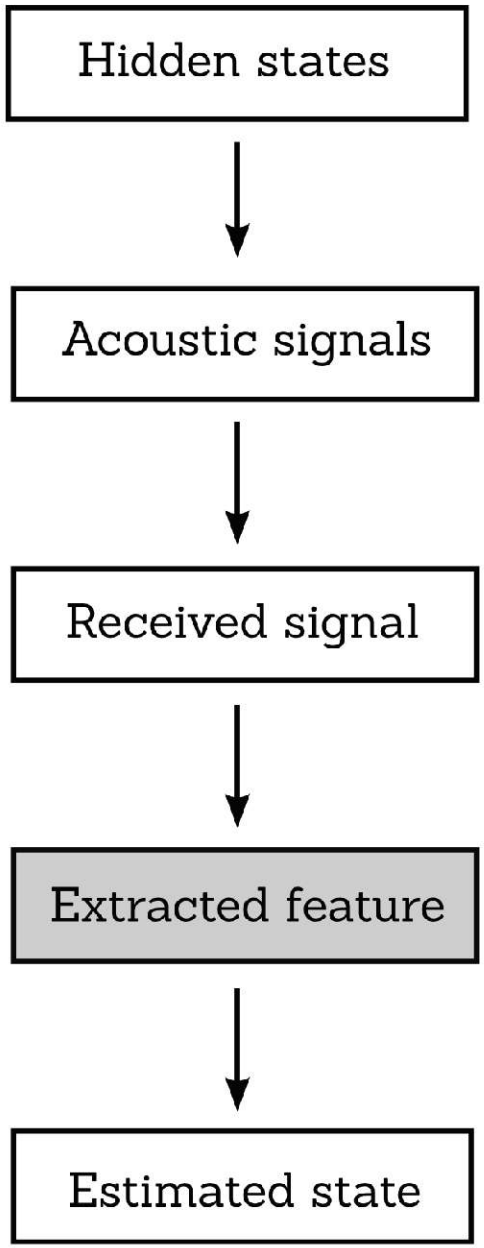
ASA

- What's next? Hearing research...

- Focus: modeling of information processing.

- At this stage: data transformation/ compression that results in redundancy reduction

- Called: feature extraction



CASA:

I. 3. MODELING FRAMEWORK

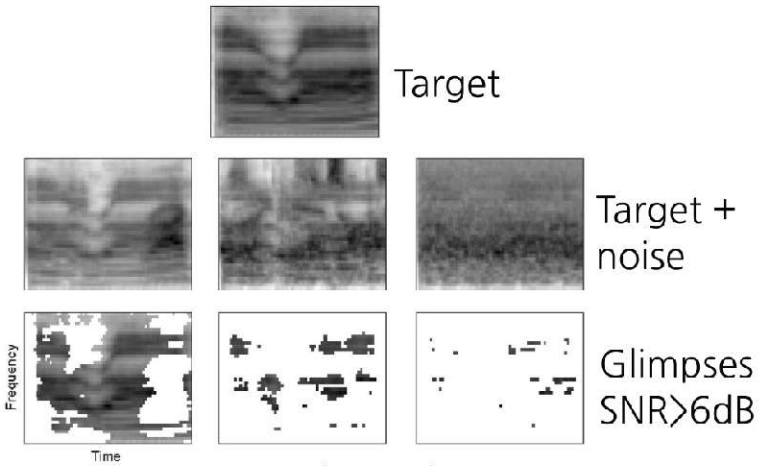
CARL VON OSSIETZKY universität OLDENBURG

ASA

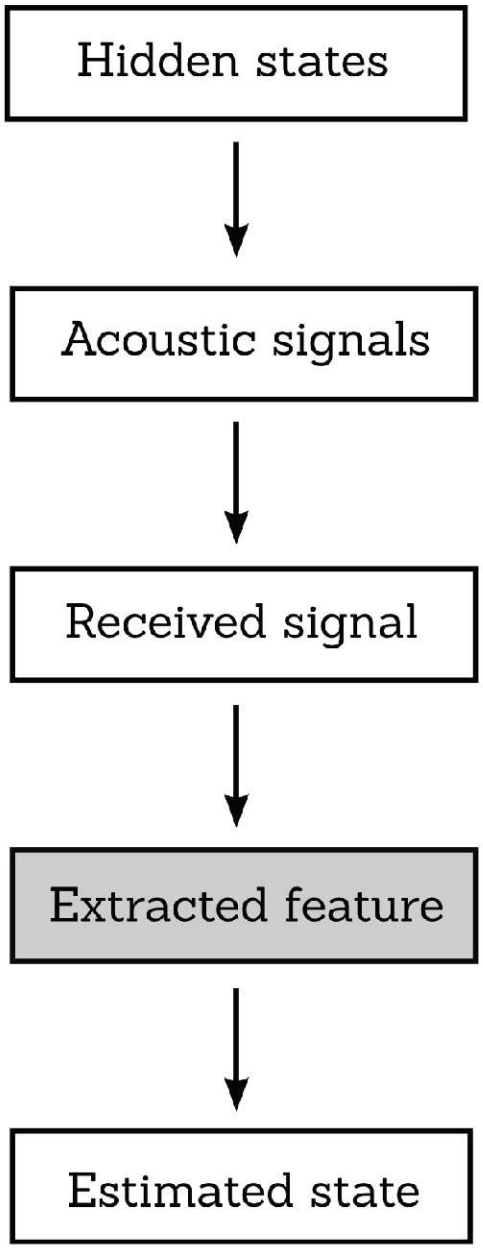
GLIMPSES

- Sparse spectro-temporal pieces of information least affected by the background.

- 'pixels' with high SNR



Cooke (2006)



CASA:

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

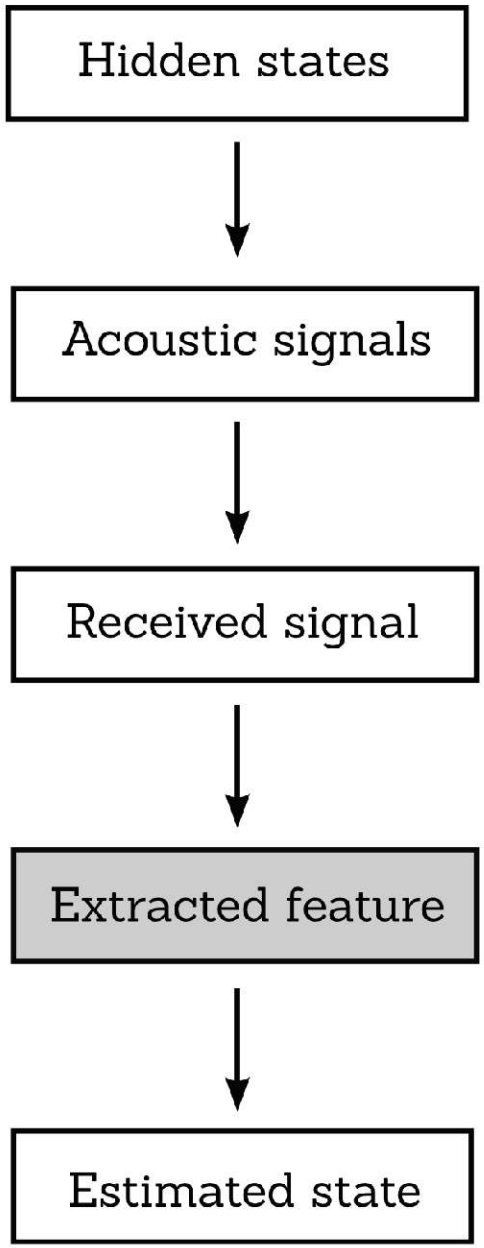
GLIMPSES

- sparse
- salient
- robust
- not a mixture

Studies* show:

Target-related glimpses are important for solving certain tasks in a complex multi-talker scene.

* Cooke (2006)
* Schoenmaker and van den Paar (2016)
* Josupeit et al. (2016)



CASA:

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

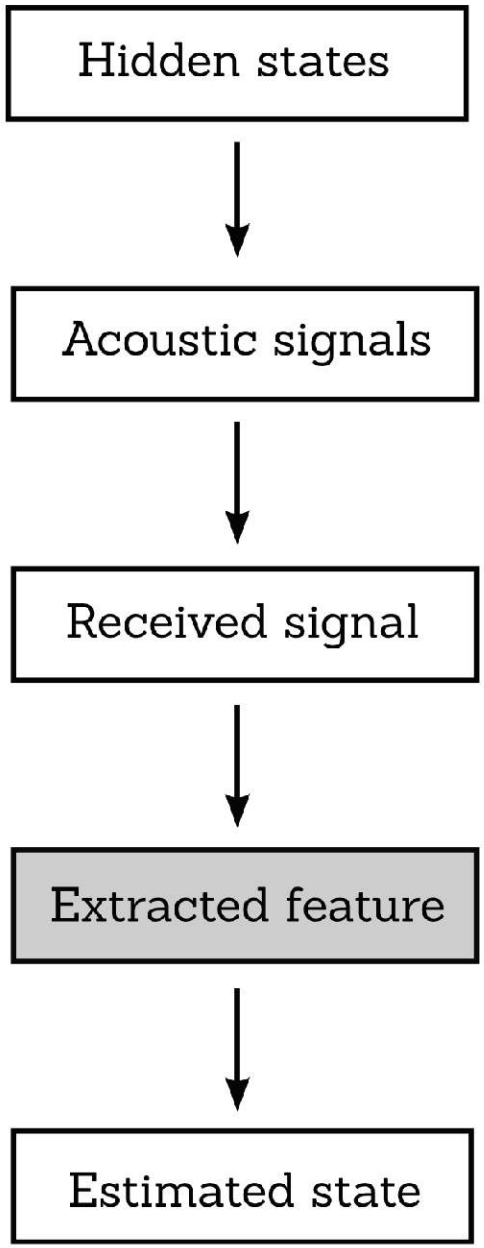
GLIMPSES

- sparse
- salient
- robust
- not a mixture

Studies* show:

Target-related glimpses are important for solving certain tasks in a complex multi-talker scene.

* Cooke (2006)
* Schoenmaker and van den Paar (2016)
* Josupeit et al. (2016)



CASA:

Problem: How to find target-related glimpses without having the perfect knowledge about the individual target and masker energies?

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

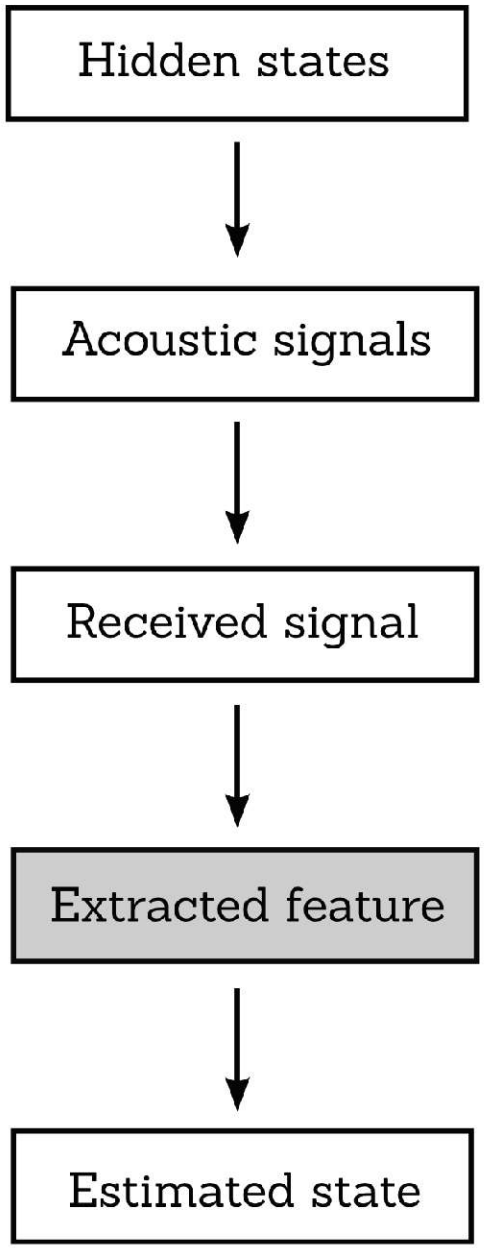
GLIMPSES

- sparse
- salient
- robust
- not a mixture

Studies* show:

Target-related glimpses are important for solving certain tasks in a complex multi-talker scene.

* Cooke (2006)
* Schoenmaker and van den Paar (2016)
* Josupeit et al. (2016)



CASA:

PERIODICITY-BASED GLIMPSING FEATURES

Josupeit and Hohmann (2017)

high periodic energy indicates a robust piece of information coming from speech

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

GLIMPSES

- sparse
- salient
- robust
- not a mixture

Studies* show:

Target-related glimpses are important for solving certain tasks in a complex multi-talker scene.

* Cooke (2006)
* Schoenmaker and van den Paar (2016)
* Josupeit et al. (2016)

Hidden states

↓

Acoustic signals

↓

Received signal

↓

Extracted feature

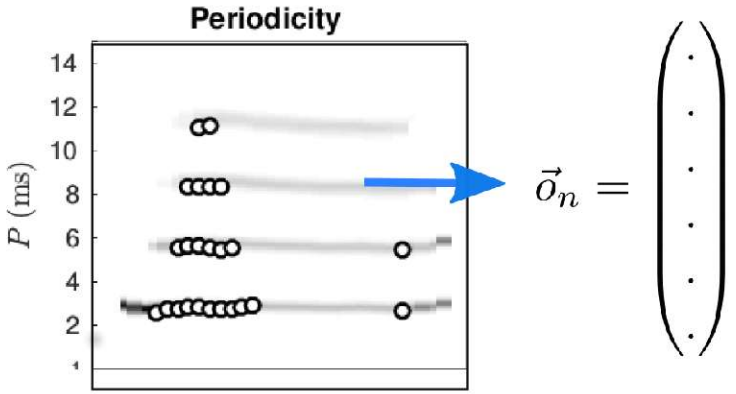
↓

Estimated state

CASA:

PERIODICITY-BASED GLIMPSING FEATURES

Josupeit and Hohmann (2017)

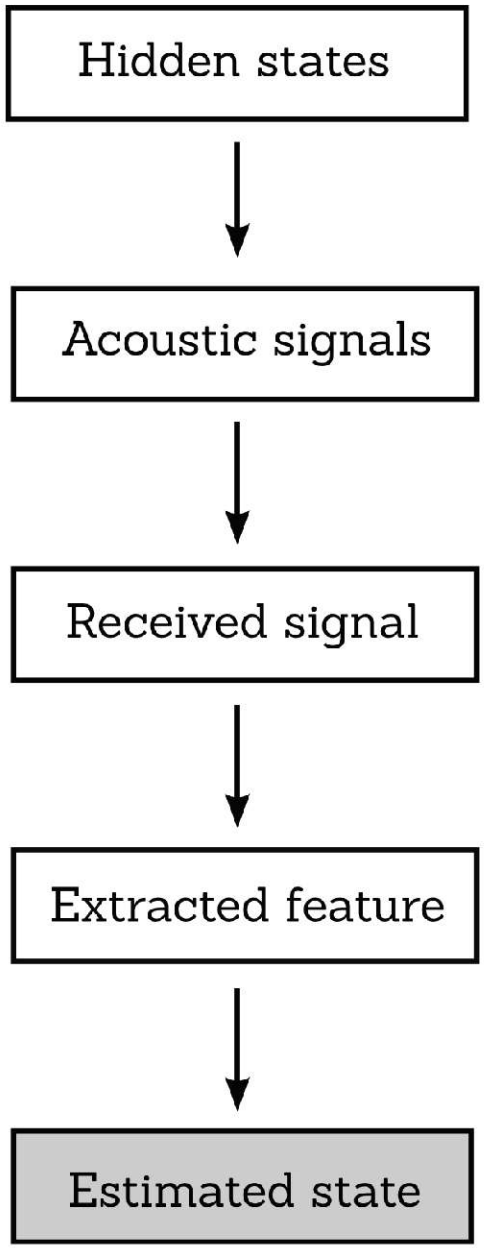


multidimensional **observation vector**

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA



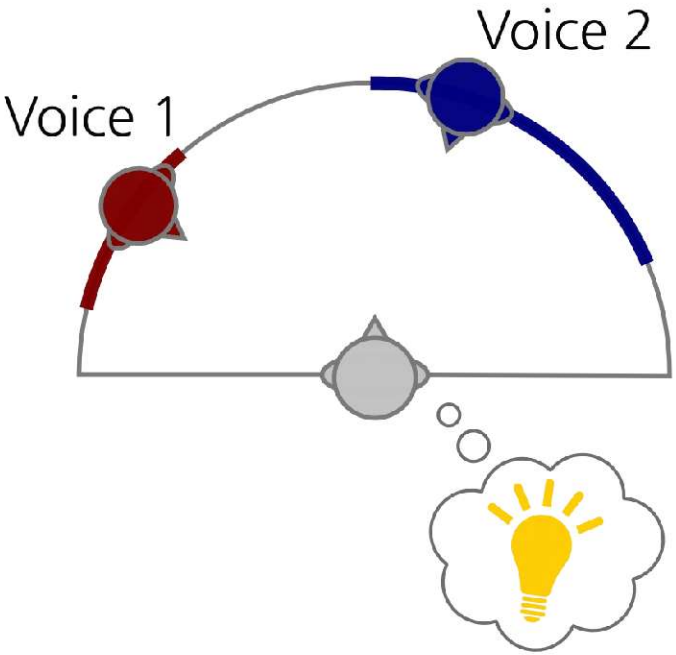
CASA:

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

- conclusions about the reality



INFERRED STATE OF A SYSTEM: description of what we think is happening

Hidden states

↓

Acoustic signals

↓

Received signal

↓

Extracted feature

↓

Estimated state

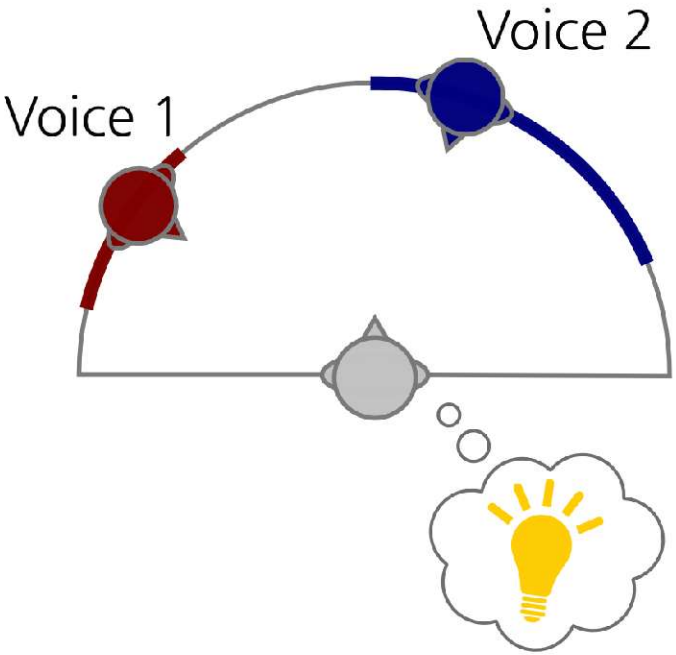
CASA:

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

- conclusions about the reality



INFERRED STATE OF A SYSTEM: description of what we think is happening

Hidden states

↓

Acoustic signals

↓

Received signal

↓

Extracted feature

↓

Estimated state

CASA:

multidimensional state vectors

$$\hat{\vec{s}}_{1,n} = \begin{pmatrix} \hat{F}0_{1,n} \\ \hat{F}1_{1,n} \\ \hat{F}2_{1,n} \\ \hat{\alpha}_{1,n} \end{pmatrix}$$ Voice 1

$$\hat{\vec{s}}_{2,n} = \begin{pmatrix} \hat{F}0_{2,n} \\ \hat{F}1_{2,n} \\ \hat{F}2_{2,n} \\ \hat{\alpha}_{2,n} \end{pmatrix}$$ Voice 2

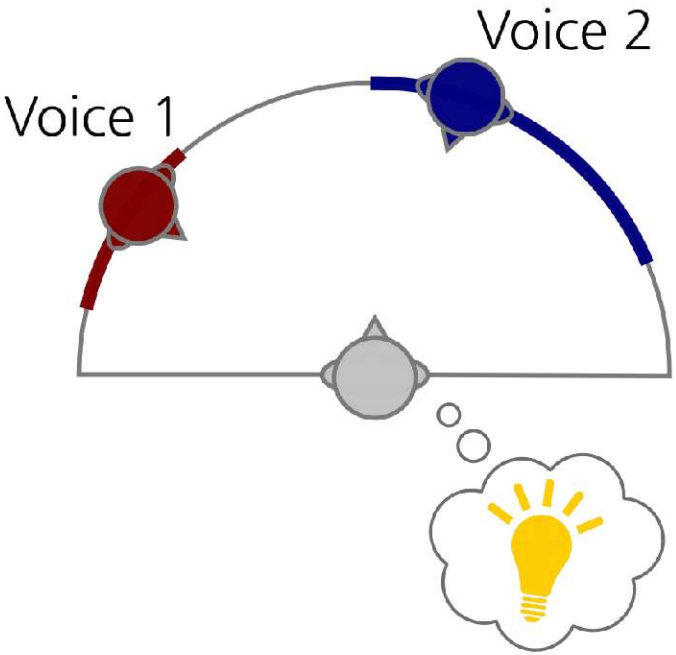
ESTIMATED STATE

I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

- conclusions about the reality



INFERRED STATE OF A SYSTEM:
description of what we think is happening

Hidden states

↓

Acoustic signals

↓

Received signal

↓

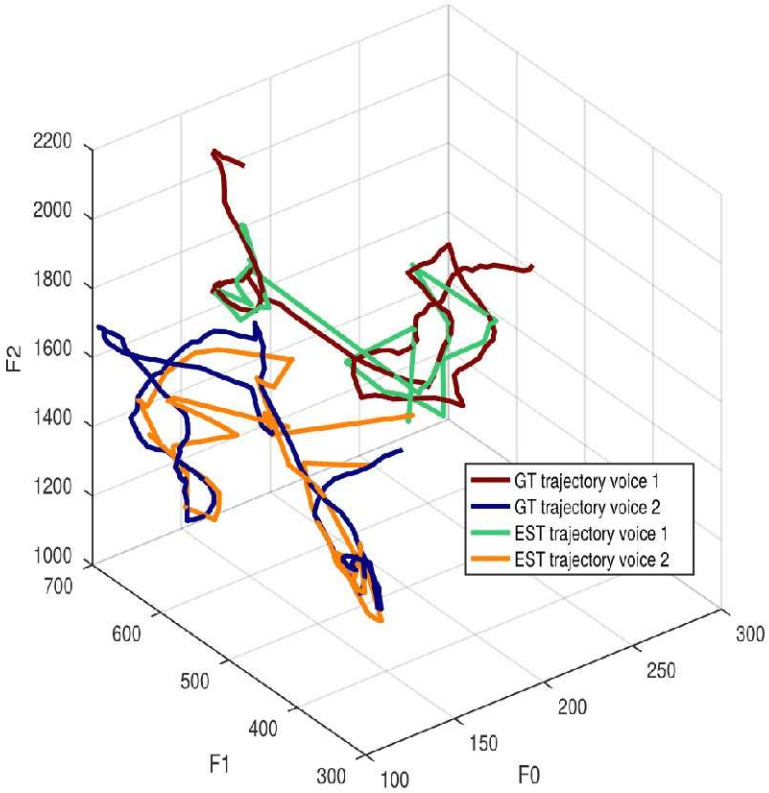
Extracted feature

↓

Estimated state

CASA:

state in time :
state trajectories



ESTIMATED STATE TRAJECTORY

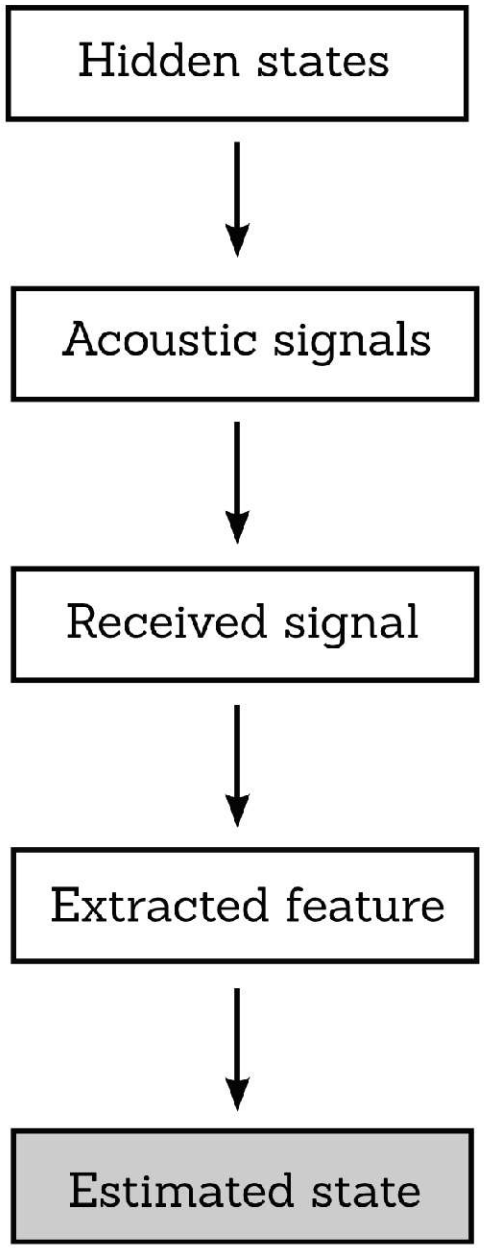
I. 3. MODELING FRAMEWORK

CARL VON OSSIETZKY universität OLDENBURG

ASA

- But how is the state estimation done?

- **Theory**: Inference made by constantly creating and comparing hypotheses about the state of the system with the sensory input.



CASA:

I. 3. MODELING FRAMEWORK

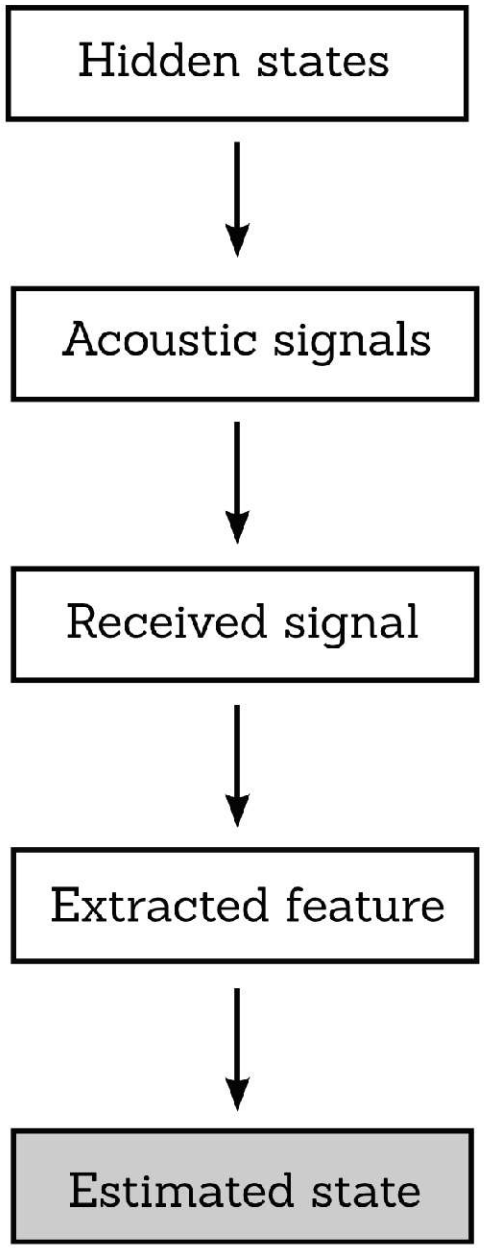
CARL VON OSSIETZKY universität OLDENBURG

ASA

PREDICTIVE CODING

- But how is the state estimation done?

- **Theory**: Inference made by constantly creating and comparing hypotheses about the state of the system with the sensory input.



CASA:

I. 3. MODELING FRAMEWORK

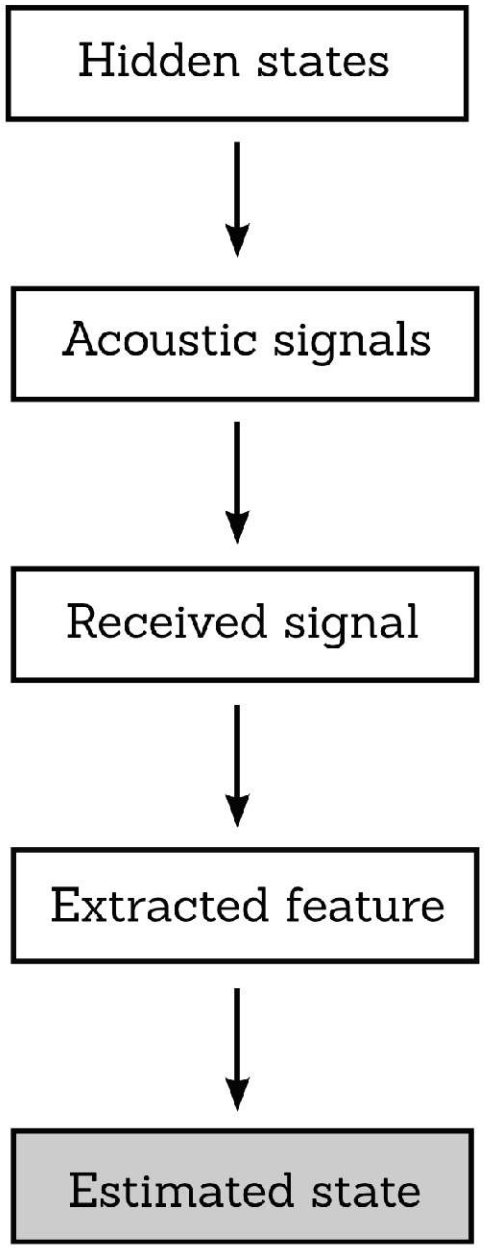
CARL VON OSSIETZKY universität OLDENBURG

ASA

PREDICTIVE CODING

- But how is the state estimation done?

- **Theory**: Inference made by constantly creating and comparing hypotheses about the state of the system with the sensory input.



CASA:

Math form of sequential competing hypotheses:

$$p(\vec{s}_n|\vec{o}_{0:n}) = \frac{p(\vec{o}_n|\vec{s}_n)p(\vec{s}_n|\vec{o}_{0:n-1})}{p(\vec{o}_n|\vec{o}_{0:n-1})}$$

SEQUENTIAL BAYESIAN ESTIMATION

- Given a sequence of observations up to now, estimate the **current state distribution**.

- Having this distribution, estimate the **most likeli state** of the system.

I. 3. MODELING FRAMEWORK

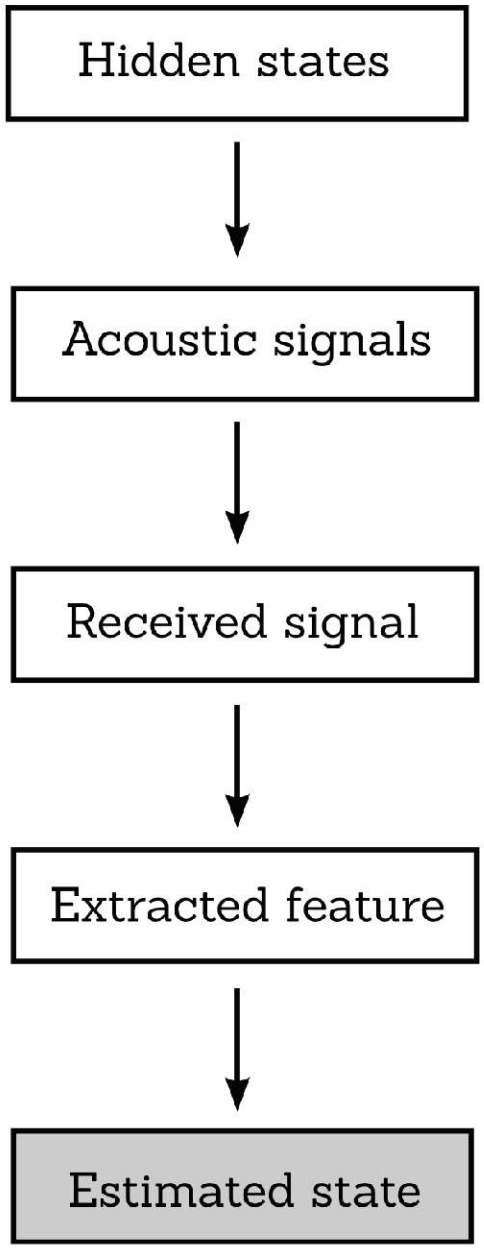
CARL VON OSSIETZKY universität OLDENBURG

ASA

PREDICTIVE CODING

- But how is the state estimation done?

- **Theory**: Inference made by constantly creating and comparing hypotheses about the state of the system with the sensory input.



CASA:

Math form of sequential competing hypotheses:

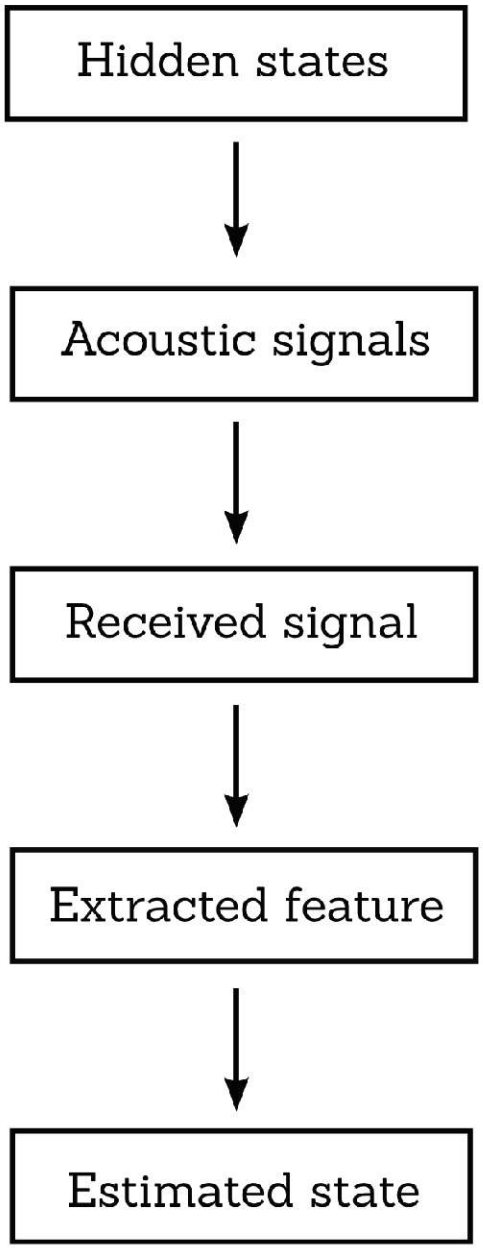
$$p(\vec{s}_n|\vec{o}_{0:n}) = \frac{p(\vec{o}_n|\vec{s}_n)p(\vec{s}_n|\vec{o}_{0:n-1})}{p(\vec{o}_n|\vec{o}_{0:n-1})}$$

SEQUENTIAL BAYESIAN ESTIMATION

No analytical solution.
Approximation:

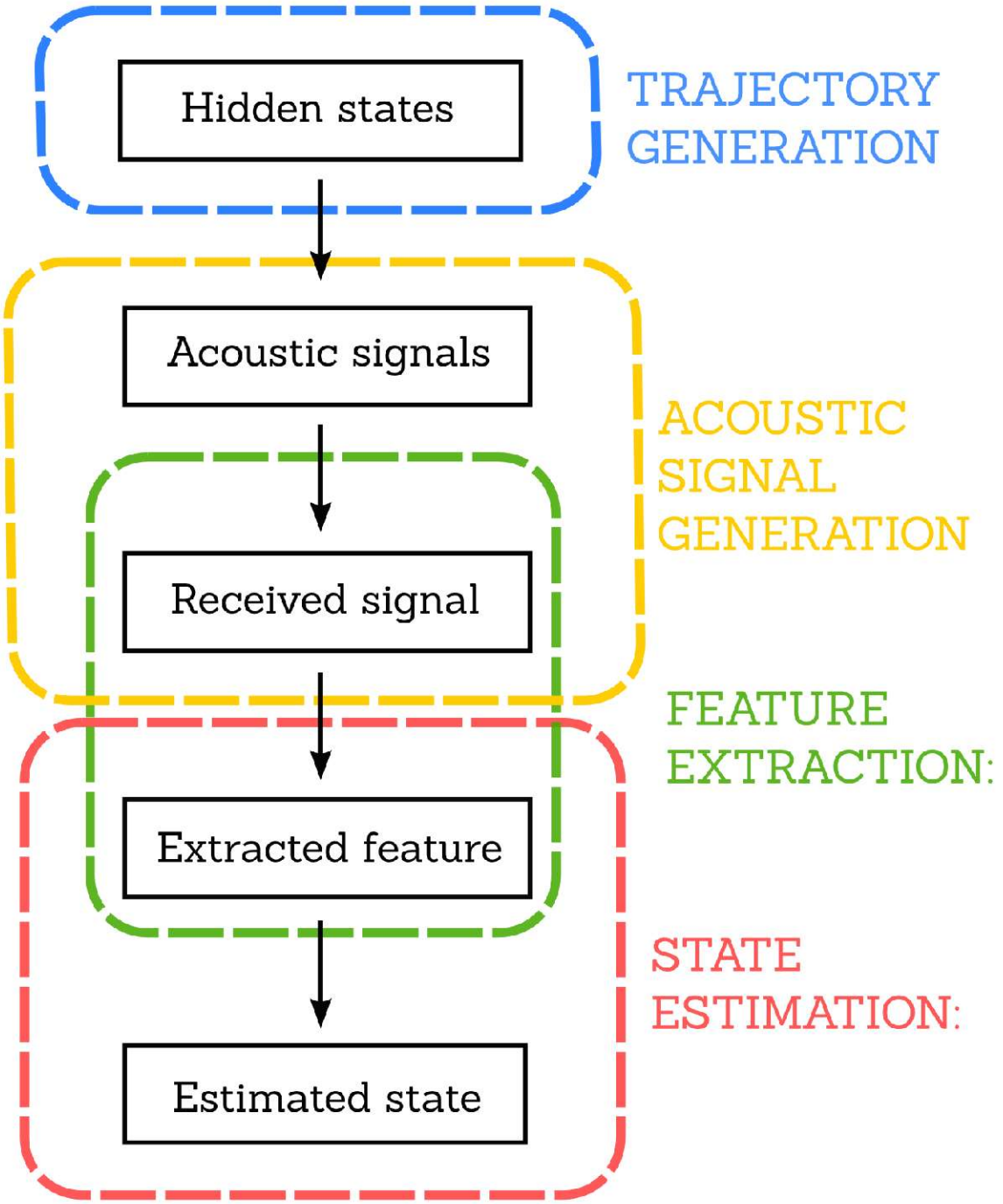
PARTICLE FILTERING

I. 3. MODELING FRAMEWORK



I. 3. MODELING FRAMEWORK

How to approach it?



I. INTRODUCTION

I. 4. SIMULATION APPROACH