The importance of auditory illusions for artificial listeners

Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

Outline

1 Computational Auditory Scene Analysis
2 A survey of CASA
3 Illusions & prediction-driven CASA
4 CASA and speech recognition
5 Implications for duplex perception
6 Conclusions

Computational Auditory Scene Analysis: An overview and some observations

Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

Outline

1 Modeling Auditory Scene Analysis
2 A survey of CASA
3 Prediction-driven CASA
4 CASA and speech recognition
5 Implications for other domains
6 Conclusions

1 Auditory Scene Analysis

"The organization of complex sound scenes according to their inferred sources"

- **Sounds rarely occur in isolation**
  - getting useful information from real-world sound requires auditory organization
- **Human audition is very effective**
  - unexpectedly difficult to model
- **'Correct' analysis defined by goal**
  - human beings have particular interests...
  - (in)dependence as the key attribute of a source
  - ecological constraints enable organization

Computational Auditory Scene Analysis (CASA)

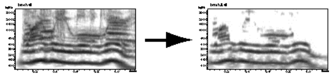
- **Automatic sound organization?**
  - convert an undifferentiated signal into a description in terms of different sources
- **Psychoacoustics defines grouping 'rules'**
  - e.g. [Bregman 1990]
  - translate into computer programs?
- **Motivations & Applications**
  - it's a puzzle: new processing principles?
  - real-world interactive systems (speech, robots)
  - hearing prostheses (enhancement, description)
  - advanced processing (remixing)
  - multimedia indexing (movies etc.)

2 CASA survey

- **Early work on co-channel speech**
  - listeners benefit from pitch difference
  - algorithms for separating periodicities
- **Utterance-sized signals need more**
  - cannot predict number of signals (0, 1, 2 ...)
  - birth/death processes
- **Ultimately, more constraints needed**
  - nonperiodic signals
  - masked cues
  - ambiguous signals

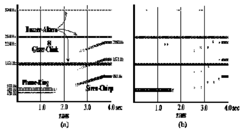
CASA1: Periodic pieces

- **Weintraub 1985**
  - separate male & female voices
  - find periodicities in each frequency channel by auto-coincidence
  - number of voices is 'hidden state'
- **Cooke & Brown (1991-3)**
  - divide time-frequency plane into elements
  - apply grouping rules to form sources
  - pull single periodic target out of noise



CASA2: Hypothesis systems

- **Okuno et al. (1994-)**
  - 'tracers' follow each harmonic + noise 'agent'
  - residue-driven: account for whole signal
- **Klassner 1996**
  - search for a combination of templates
  - high-level hypotheses permit front-end tuning



- **Ellis 1996**
  - model for events perceived in dense scenes
  - prediction-driven: observations - hypotheses

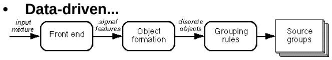
CASA3: Other approaches

- **Blind source separation (Bell & Sejnowski)**
  - find exact separation parameters by maximizing statistic e.g. signal independence
- **HMM decomposition (RK Moore)**
  - recover combined source states directly
- **Neural models (Malsburg, Wang & Brown)**
  - avoid implausible AI methods (search, lists)
  - oscillators substitute for iteration?

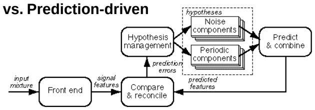
3 Prediction-driven CASA

Perception is not *direct*
but a *search* for *plausible hypotheses*

- **Data-driven...**



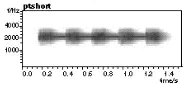
vs. Prediction-driven



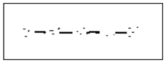
- **Novel features**
  - reconcile complete explanation to input
  - 'vocabulary' of noise/transient/periodic
  - multiple hypotheses
  - sufficient detail for reconstruction
  - explanation hierarchy

Analyzing the continuity illusion

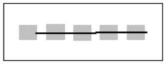
- **Interrupted tone heard as continuous**
  - .. if the interruption could be a masker



- **Data-driven just sees gaps**

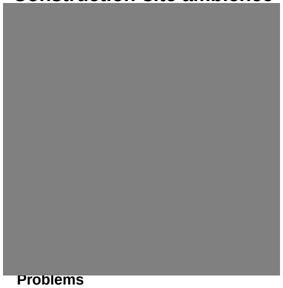


- **Prediction-driven can accommodate**



  - special case or general principle?

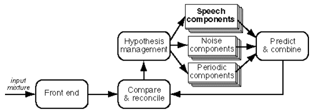
PDCASA example: Construction-site ambience



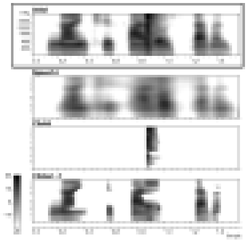
- **Problems**
  - error allocation
  - source hierarchy
  - rating hypotheses
  - resynthesis

4 CASA for speech recognition

- **Speech recognition is very fragile**
  - lots of motivation to use ‘source separation’
- **Recognize combined states? (Moore)**
  - ‘state’ becomes very complex
- **Data-driven: CASA as preprocessor**
  - problems with ‘holes’ (Cooke, Okuno)
  - doesn’t exploit knowledge of speech structure
- **Prediction-driven: speech as component**
  - same ‘reconciliation’ of speech hypotheses
  - need to express ‘predictions’ in signal domain



Example of speech & nonspeech



- **Problems:**
  - undoing classification & normalization
  - finding a starting hypothesis
  - granularity of integration

5 Prediction-driven analysis and duplex perception

- **Single element → 2 percepts?**
  - e.g. contralateral formant transition
  - doesn’t fit into exclusive support hierarchy
- **But: two elements at same position**
  - hypotheses suggest overlap
  - predictions combine
  - reconciliation is OK
- **Order debate is sidestepped**
  - .. not a left-to-right data path

Duplex perception as masking & restoration

- **Account for masking could ‘work’ for duplex**
  - bilateral masking levels?
  - masking spread?
  - tolerable colorations?
- **Sinewave speech as a plausible masker?**
  - formants hiding under each whistle?
  - greedy speech hypothesis generator
- **Problems:**
  - where do hypotheses come from? (priming)
  - what limits on illusory speech?

5 Lessons for other domains

- **Problem: inadequate signal data**
  - hearing: masking
  - vision: occlusion
  - other sensor domains: noise/limits
- **General answer: employ constraints**
  - high-level prior expectations
  - mid-level regularities
  - low-level continuity
- **Hearing is a admirable solution**
- **Prediction-driven approach suggests priorities**

Essential features of PDCASA

- **Prediction-reconciliation of hypotheses**
  - specific hypotheses are pursued
  - lack-of-refutation standard
- **Provide a complete explanation**
  - keeping track of the obstruction can help in compensating for its effects
- **Hierarchic representation**
  - useful constraints occur at many levels: want to be able to apply where appropriate
- **Preserve detail**
  - even when resynthesis is not a goal
  - helps gauge goodness-of-fit

6 Conclusions

- **Auditory organization is indispensable in real environments**
- **We don’t know how listeners do it!**
  - plenty of modeling interest
- **Prediction-reconciliation can account for ‘illusions’**
  - use ‘knowledge’ when signal is inadequate
  - important in a wider range of circumstances?
- **Speech recognizers are a good source of knowledge**
- **Wider implications of the prediction-driven approach**
  - understanding perceptual paradoxes
  - applications in other domains

The importance of auditory illusions for artificial listeners

Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

Outline

1. **Computational Auditory Scene Analysis**
2. **A survey of CASA**
3. **Illusions & prediction-driven CASA**
4. **CASA and speech recognition**
5. **Implications for duplex perception**
6. **Conclusions**

Computational Auditory Scene Analysis: An overview and some observations

Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

Outline

1. **Modeling Auditory Scene Analysis**
2. **A survey of CASA**
3. **Prediction-driven CASA**
4. **CASA and speech recognition**
5. **Implications for other domains**
6. **Conclusions**

1

Auditory Scene Analysis

“The organization of complex sound scenes according to their inferred sources”

- **Sounds rarely occur in isolation**
  - getting useful information from real-world sound requires auditory organization
- **Human audition is very effective**
  - unexpectedly difficult to model
- **‘Correct’ analysis defined by goal**
  - human beings have particular interests...
  - (in)dependence as the key attribute of a source
  - ecological constraints enable organization

Computational Auditory Scene Analysis (CASA)

- **Automatic sound organization?**
  - convert an undifferentiated signal into a description in terms of different sources
- **Psychoacoustics defines grouping ‘rules’**
  - e.g. [Bregman 1990]
  - translate into computer programs?
- **Motivations & Applications**
  - it’s a puzzle: new processing principles?
  - real-world interactive systems (speech, robots)
  - hearing prostheses (enhancement, description)
  - advanced processing (remixing)
  - multimedia indexing (movies etc.)

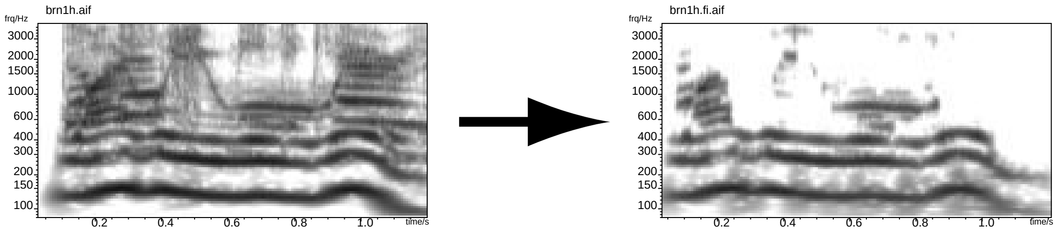
2

CASA survey

- **Early work on co-channel speech**
  - listeners benefit from pitch difference
  - algorithms for separating periodicities
- **Utterance-sized signals need more**
  - cannot predict number of signals (0, 1, 2 ...)
  - birth/death processes
- **Ultimately, more constraints needed**
  - nonperiodic signals
  - masked cues
  - ambiguous signals

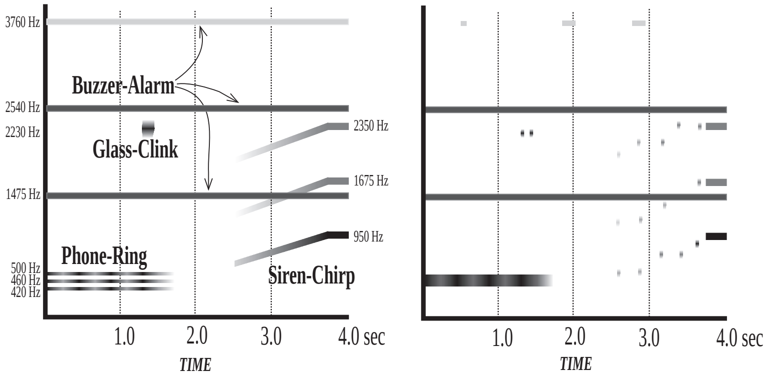
CASA1: Periodic pieces

- **Weintraub 1985**
  - separate male & female voices
  - find periodicities in each frequency channel by auto-coincidence
  - number of voices is ‘hidden state’
- **Cooke & Brown (1991-3)**
  - divide time-frequency plane into elements
  - apply grouping rules to form sources
  - pull single periodic target out of noise



CASA2: Hypothesis systems

- **Okuno et al. (1994-)**
  - ‘tracers’ follow each harmonic + noise ‘agent’
  - residue-driven: account for whole signal
- **Klassner 1996**
  - search for a combination of templates
  - high-level hypotheses permit front-end tuning

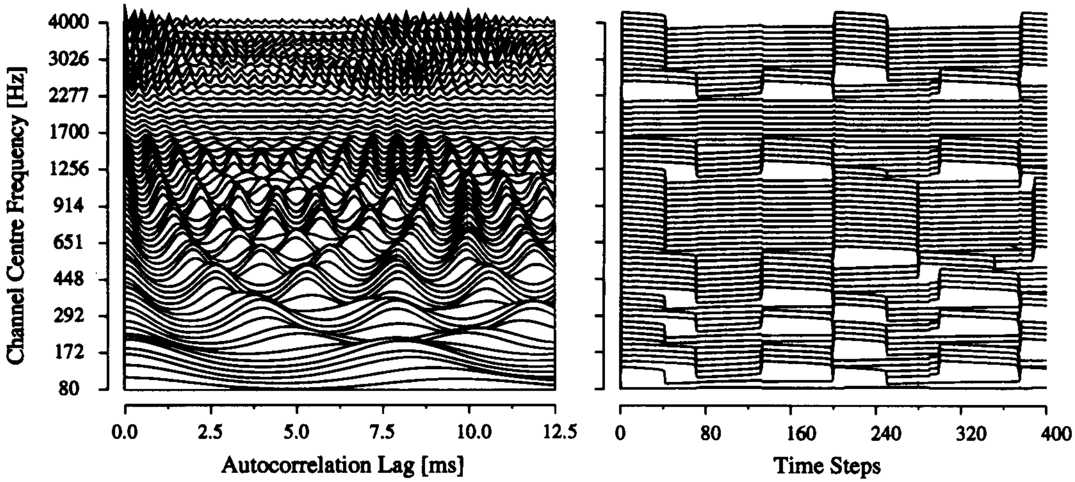


(a) (b)

- **Ellis 1996**
  - model for events perceived in dense scenes
  - prediction-driven: observations - hypotheses

CASA3: Other approaches

- **Blind source separation (Bell & Sejnowski)**
  - find exact separation parameters by maximizing statistic e.g. signal independence
- **HMM decomposition (RK Moore)**
  - recover combined source states directly
- **Neural models (Malsburg, Wang & Brown)**
  - avoid implausible AI methods (search, lists)
  - oscillators substitute for iteration?

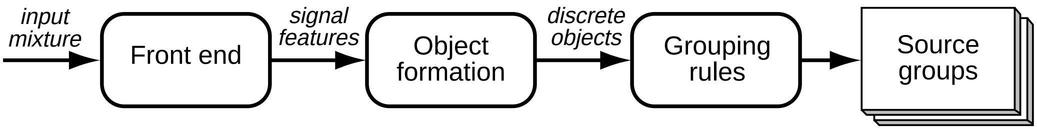


3

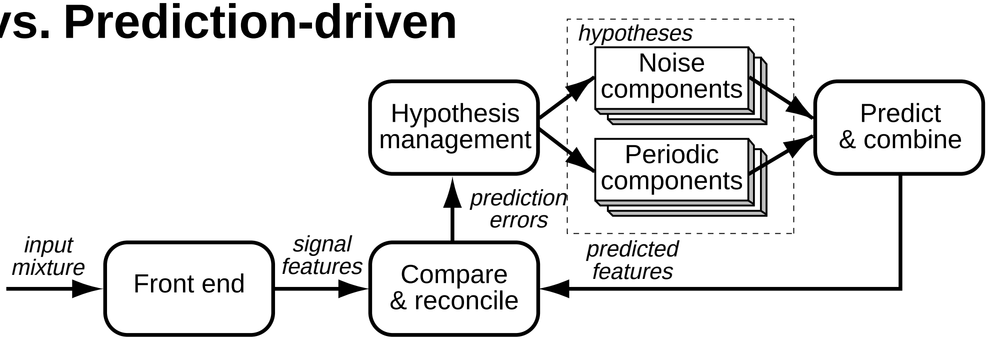
Prediction-driven CASA

Perception is not *direct*
but a *search* for *plausible hypotheses*

- **Data-driven...**



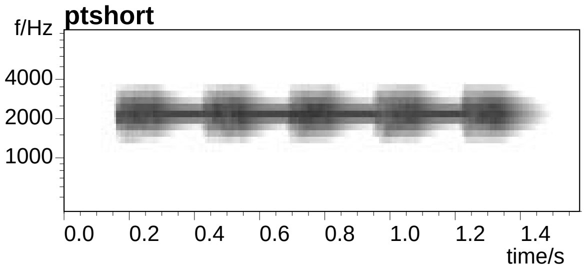
vs. Prediction-driven



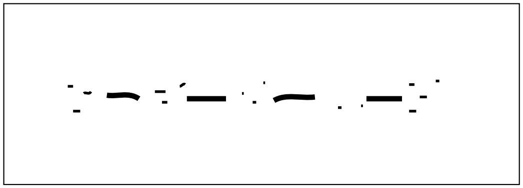
- **Novel features**
  - reconcile complete explanation to input
  - 'vocabulary' of noise/transient/periodic
  - multiple hypotheses
  - sufficient detail for reconstruction
  - explanation hierarchy

Analyzing the continuity illusion

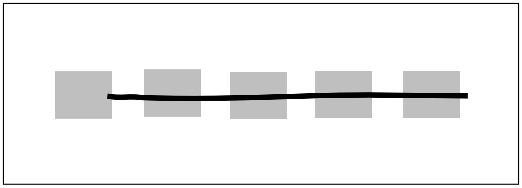
- **Interrupted tone heard as continuous**
  - .. if the interruption could be a masker



- **Data-driven just sees gaps**

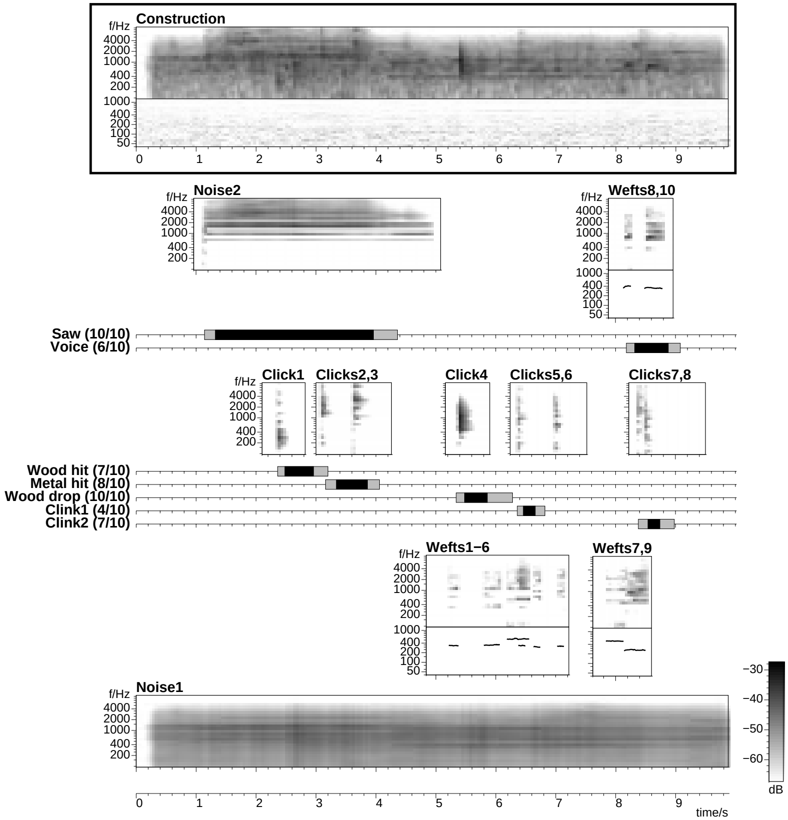


- **Prediction-driven can accommodate**



  - special case or general principle?

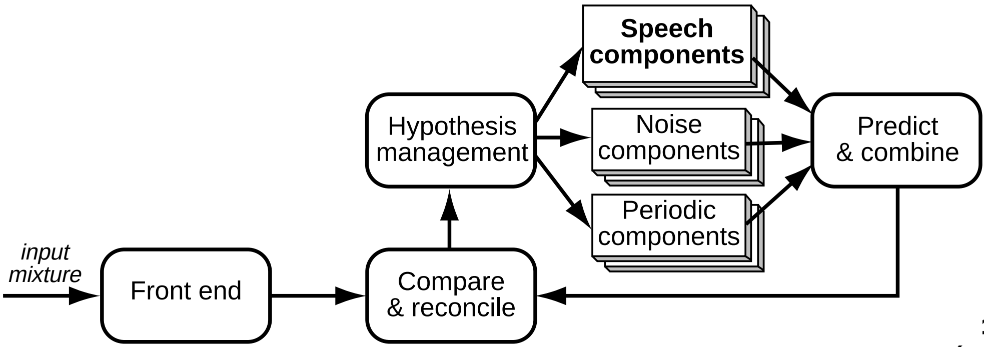
PDCASA example: Construction-site ambience



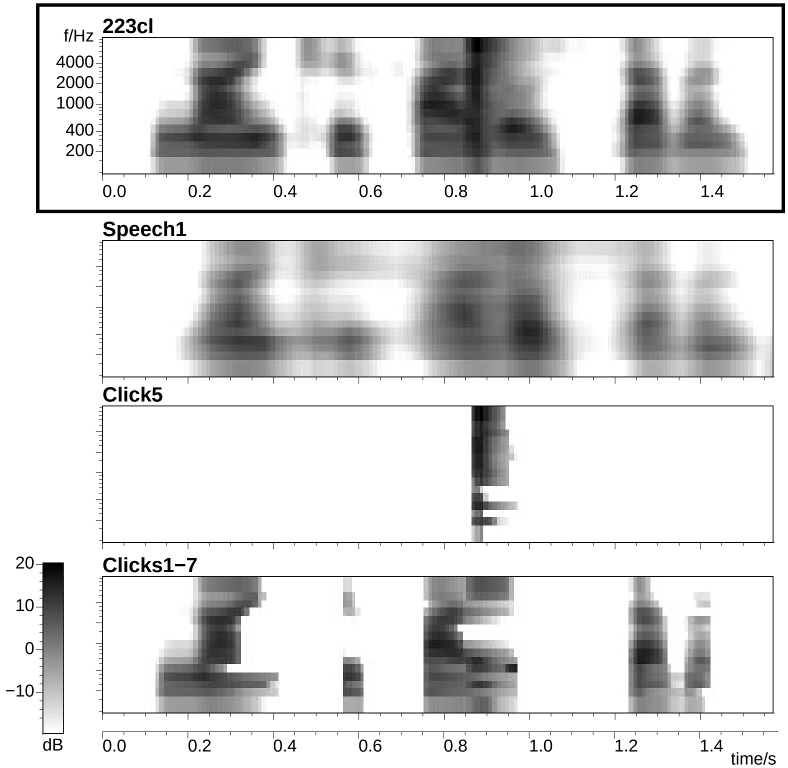
- **Problems**
  - error allocation
  - source hierarchy
  - rating hypotheses
  - resynthesis

4 CASA for speech recognition

- **Speech recognition is very fragile**
  - lots of motivation to use 'source separation'
- **Recognize combined states? (Moore)**
  - 'state' becomes very complex
- **Data-driven: CASA as preprocessor**
  - problems with 'holes' (Cooke, Okuno)
  - doesn't exploit knowledge of speech structure
- **Prediction-driven: speech as component**
  - same 'reconciliation' of speech hypotheses
  - need to express 'predictions' in signal domain



Example of speech & nonspeech



- **Problems:**
  - undoing classification & normalization
  - finding a starting hypothesis
  - granularity of integration

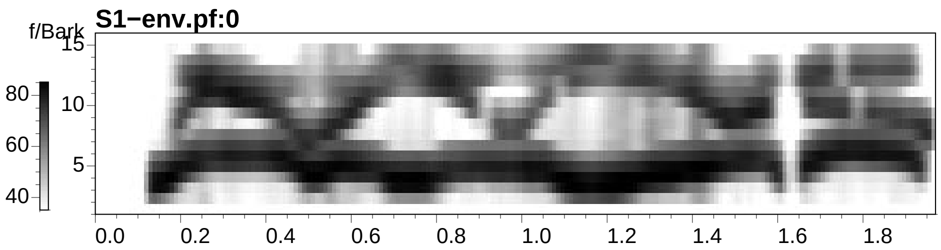
5

Prediction-driven analysis and duplex perception

- **Single element → 2 percepts?**
  - e.g. contralateral formant transition
  - doesn’t fit into exclusive support hierarchy
- **But: two elements at same position**
  - hypotheses suggest overlap
  - predictions combine
  - reconciliation is OK
- **Order debate is sidestepped**
  - .. not a left-to-right data path

Duplex perception as masking & restoration

- **Account for masking could ‘work’ for duplex**
  - bilateral masking levels?
  - masking spread?
  - tolerable colorations?
- **Sinewave speech as a plausible masker?**
  - formants hiding under each whistle?
  - greedy speech hypothesis generator
- **Problems:**
  - where do hypotheses come from? (priming)
  - what limits on illusory speech?



5 Lessons for other domains

- **Problem: inadequate signal data**
  - hearing: masking
  - vision: occlusion
  - other sensor domains: noise/limits
- **General answer: employ constraints**
  - high-level prior expectations
  - mid-level regularities
  - low-level continuity
- **Hearing is a admirable solution**
- **Prediction-driven approach suggests priorities**

Essential features of PDCASA

- **Prediction-reconciliation of hypotheses**
  - specific hypotheses are pursued
  - lack-of-refutation standard
- **Provide a complete explanation**
  - keeping track of the obstruction can help in compensating for its effects
- **Hierarchic representation**
  - useful constraints occur at many levels: want to be able to apply where appropriate
- **Preserve detail**
  - even when resynthesis is not a goal
  - helps gauge goodness-of-fit

6

Conclusions

- **Auditory organization is indispensable in real environments**
- **We don’t know how listeners do it!**
  - plenty of modeling interest
- **Prediction-reconciliation can account for ‘illusions’**
  - use ‘knowledge’ when signal is inadequate
  - important in a wider range of circumstances?
- **Speech recognizers are a good source of knowledge**
- **Wider implications of the prediction-driven approach**
  - understanding perceptual paradoxes
  - applications in other domains