Computational Auditory Scene Analysis exploiting Speech Recognition knowledge

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

Outline

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

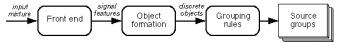
Computational Auditory Scene Analysis (CASA)

- **Automatic sound organization is desirable**
  - real-world interactive systems (speech, robots)
  - hearing prostheses (enhancement, description)
  - advanced processing (remixing)
  - multimedia indexing (movies etc.)
- **Grouping 'rules' (e.g. Bregman 1990)**
  - translate into computer programs?
- **'Data-driven' approach (e.g. Brown 1992)**
  - extract features & cues
  - form elements
  - group into sources

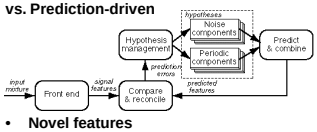
Prediction-driven CASA

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

- **Data-driven ...**



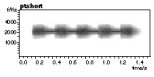
vs. Prediction-driven



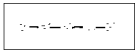
- **Novel features**
  - reconcile complete explanation to input
  - 'vocabulary' of noise/transient/periodic
  - multiple hypotheses
  - sufficient detail for reconstruction

Reproducing restoration phenomena

- **E.g. the continuity illusion**



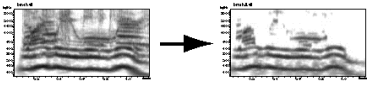
- **Data-driven approach just sees gaps**



  - (how to handle noise?)
- **Continuous tone is 'consistent' under prediction-driven approach**

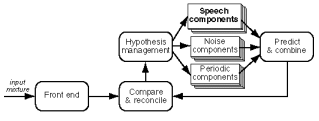
2 CASA for speech recognition

- **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' (but: Cooke, Okuno)
  - doesn't exploit knowledge of speech structure



Combining PDCASA and ASR

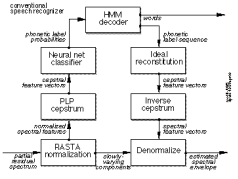
- **Prediction-driven: speech as component**



  - speech hypotheses within same reconciliation framework
  - need to express 'predictions' in signal domain
- **Each component makes a *projection* of residual**
  - into e.g. 'the space of all speech sounds'

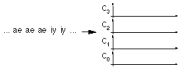
3 A speech hypothesis module

- **Speech recognition involves:**
  - normalizing & generalizing
  - classifying into phonetic state labels



- **Prediction-reconciliation requires reconstructed signal**
  - invert labels to features
  - invert features to signal

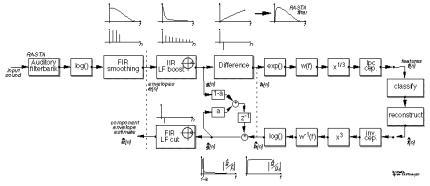
Inverting labels to features



- **Classification intrinsically many→one**
  - neural net classifier even more opaque
- **Train 'average' feature window by label**
  - i.e. just use class centers
  - overlap in reconstruction → some transition
  - .. more normalized → more representative

Inverting features to signal

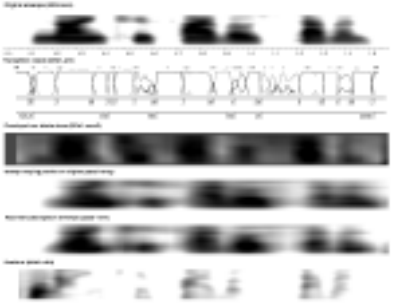
- **RASTA normalization removes average levels**
- **Solution: save slowly-varying part & restore**



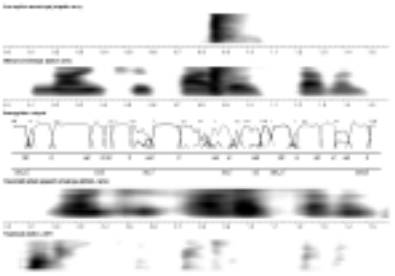
- what about generalization (blurring)?

4 Results of the modified recognizer

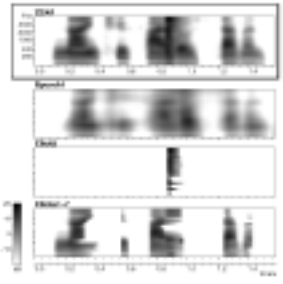
- **Reconstruct 'canonical' signal**



Example of speech & nonspeech



Putting it into the scene analyzer



- **Prediction shortfalls dominate residual**

5 Future work

- **Better signal predictions from the recognizer**
  - normalized training
  - weighted reconstruction
  - more classes?
- **Other immediate problems**
  - iteration!
  - starting hypothesis
  - granularity of integration
  - low-frequency separation
- **Nonspeech knowledge?**
- **Performance & evaluation**

Conclusions

- **Need to use scene analysis for real sounds**
- **Listeners' scene analysis relies on knowledge-based predictions**
- **Use prediction-driven formulation to employ speech-recognizer knowledge for explanation**
- **But: need better 'predictions'**
  - better inverse-classification
  - better normalization & inversion

Computational Auditory Scene Analysis exploiting Speech Recognition knowledge

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

Outline

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

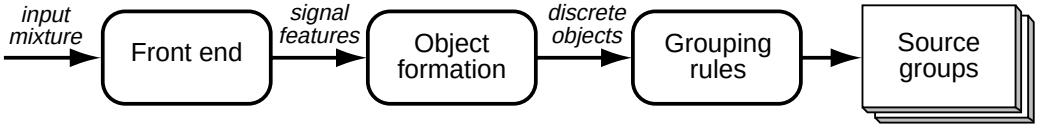
Computational Auditory Scene Analysis (CASA)

- **Automatic sound organization is desirable**
  - real-world interactive systems (speech, robots)
  - hearing prostheses (enhancement, description)
  - advanced processing (remixing)
  - multimedia indexing (movies etc.)
- **Grouping 'rules' (e.g. Bregman 1990)**
  - translate into computer programs?
- **'Data-driven' approach (e.g. Brown 1992)**
  - extract features & cues
  - form elements
  - group into sources

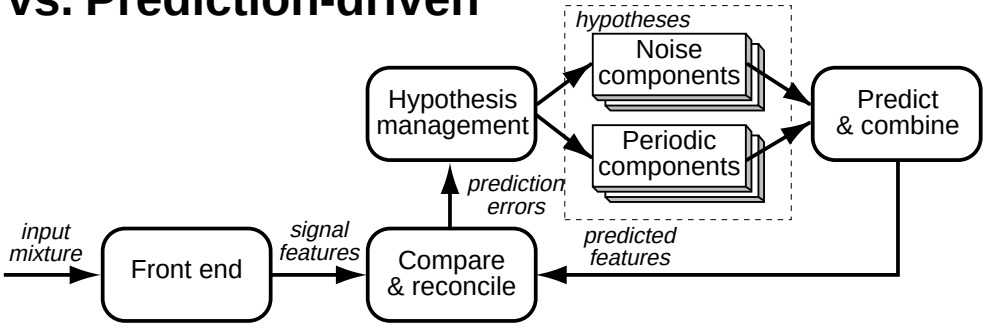
Prediction-driven CASA

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

- **Data-driven ...**



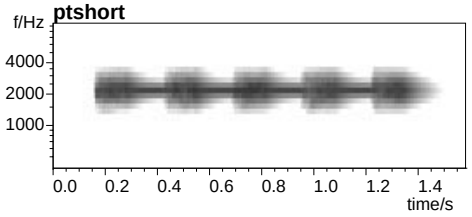
vs. Prediction-driven



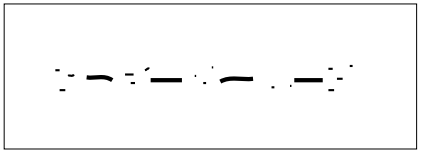
- **Novel features**
  - reconcile complete explanation to input
  - 'vocabulary' of noise/transient/periodic
  - multiple hypotheses
  - sufficient detail for reconstruction

Reproducing restoration phenomena

- **E.g. the continuity illusion**

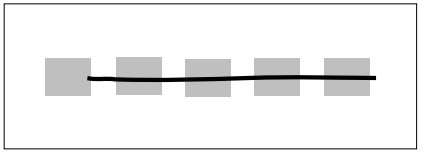


- **Data-driven approach just sees gaps**



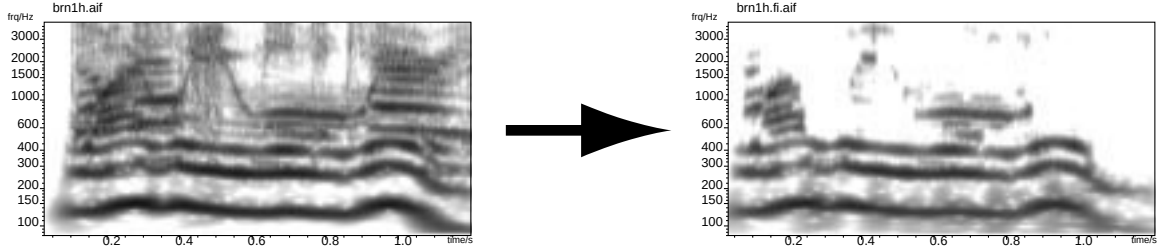
  - (how to handle noise?)

- **Continuous tone is ‘consistent’ under prediction-driven approach**



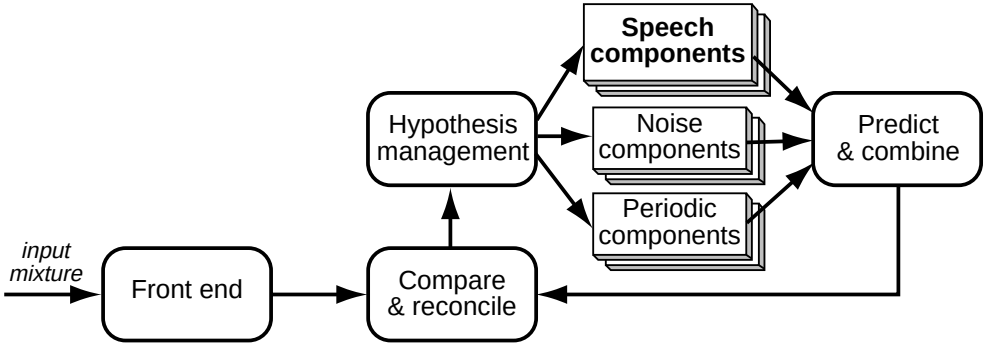
2 CASA for speech recognition

- **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’ (but: Cooke, Okuno)
  - doesn’t exploit knowledge of speech structure



Combining PDCASA and ASR

- **Prediction-driven: speech as component**

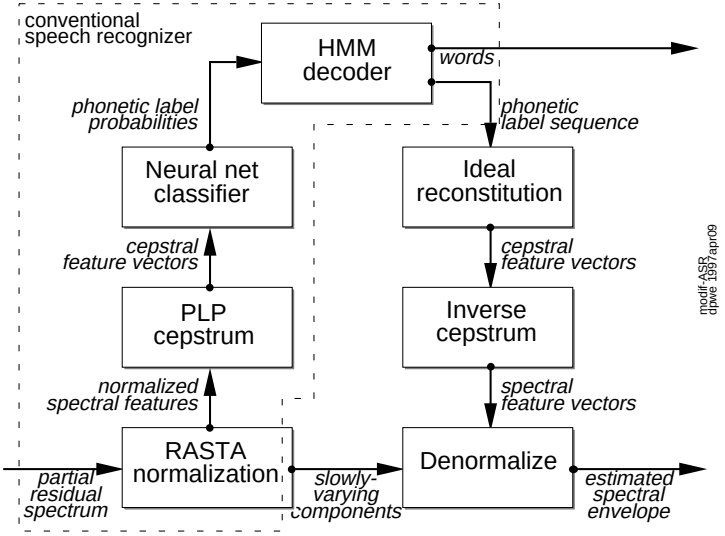


  - speech hypotheses within same reconciliation framework
  - need to express 'predictions' in signal domain

- **Each component makes a *projection* of residual**
  - into e.g. 'the space of all speech sounds'

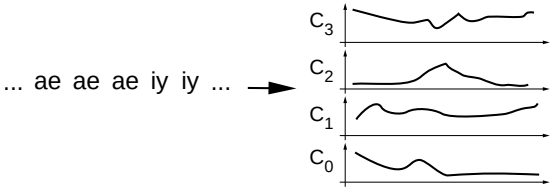
3 A speech hypothesis module

- **Speech recognition involves:**
  - normalizing & generalizing
  - classifying into phonetic state labels

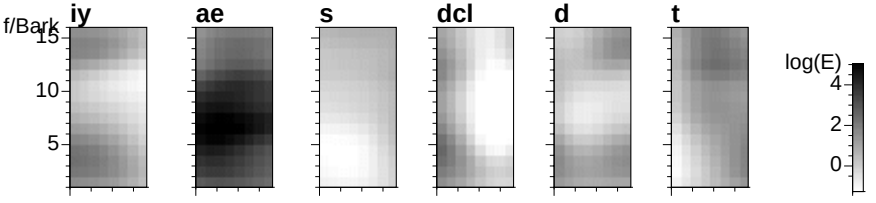


- **Prediction-reconciliation requires reconstructed signal**
  - invert labels to features
  - invert features to signal

Inverting labels to features

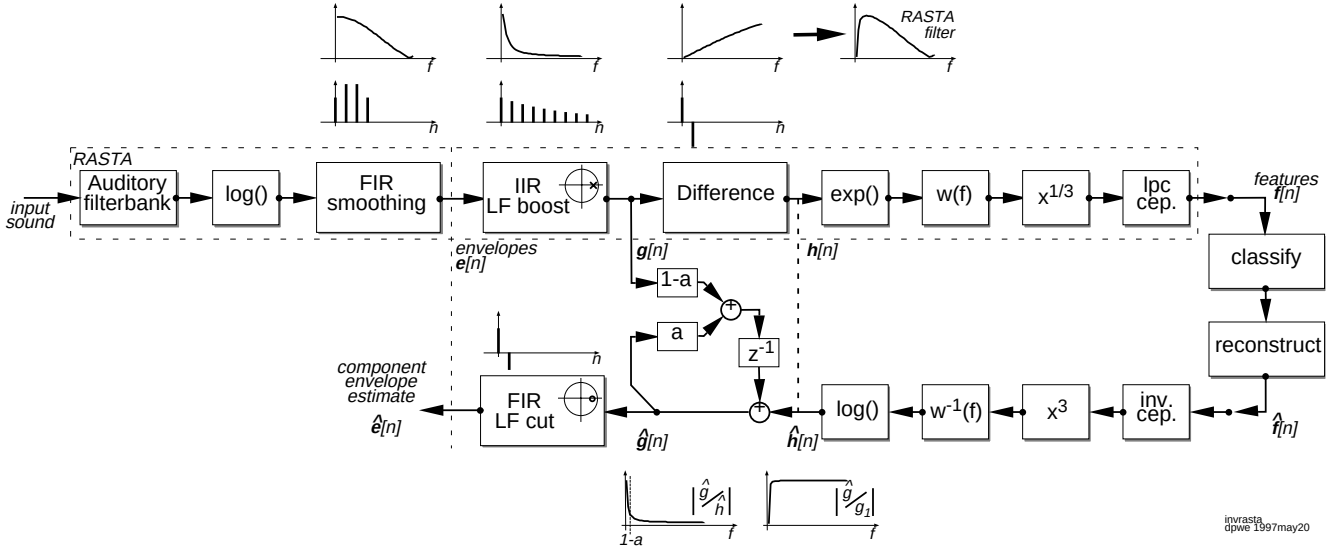


- **Classification intrinsically many→one**
  - neural net classifier even more opaque
- **Train 'average' feature window by label**
  - i.e. just use class centers
  - overlap in reconstruction → some transition
  - .. more normalized → more representative



Inverting features to signal

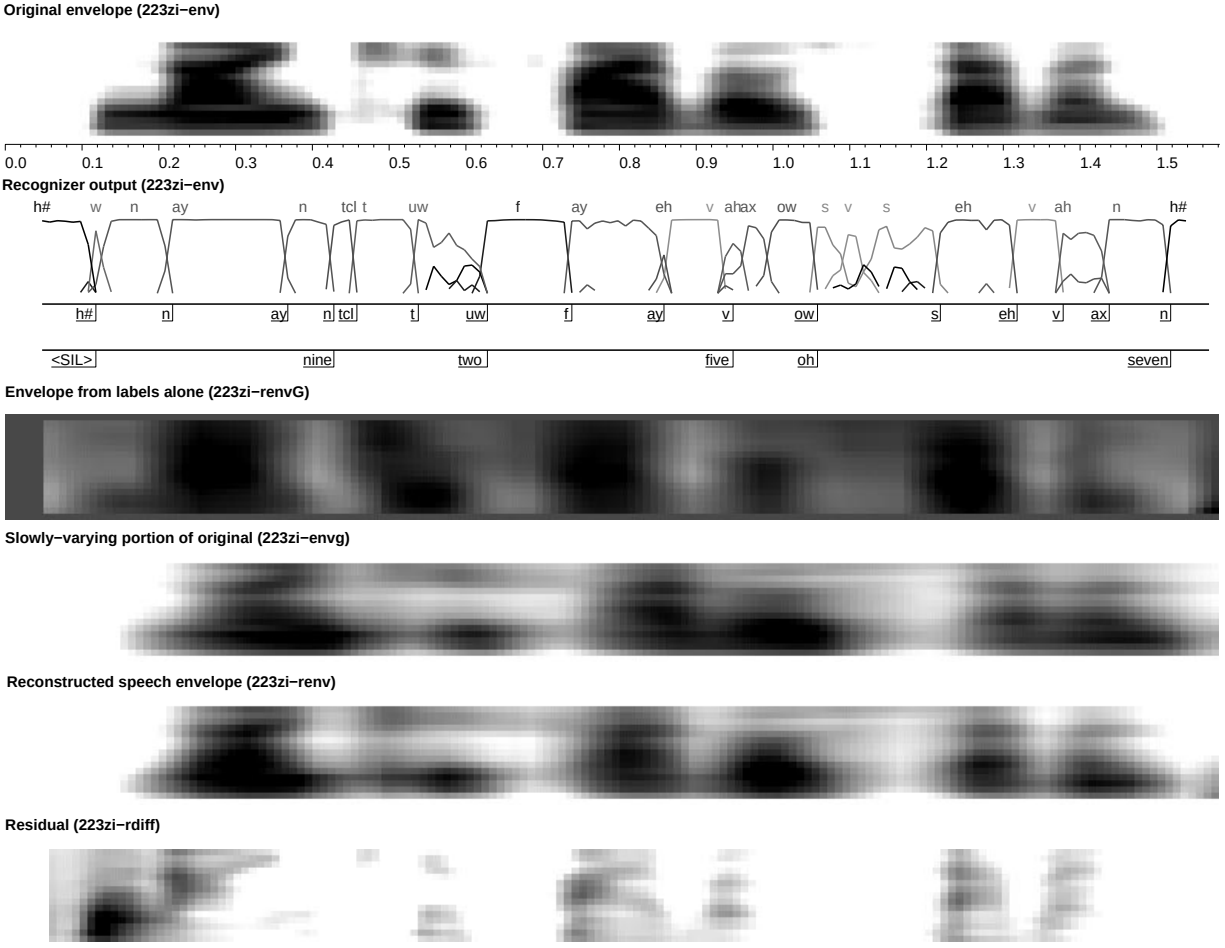
- **RASTA normalization removes average levels**
- **Solution: save slowly-varying part & restore**



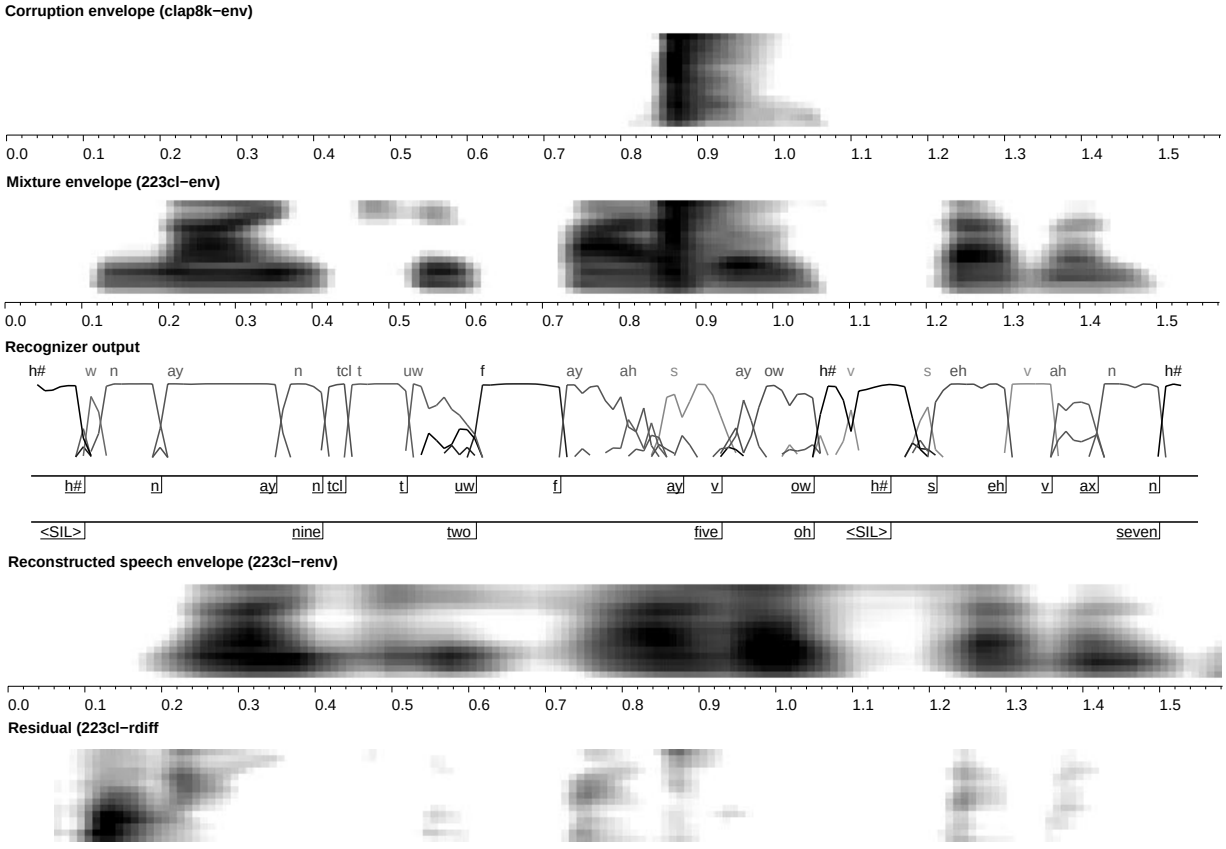
- what about generalization (blurring)?

4 Results of the modified recognizer

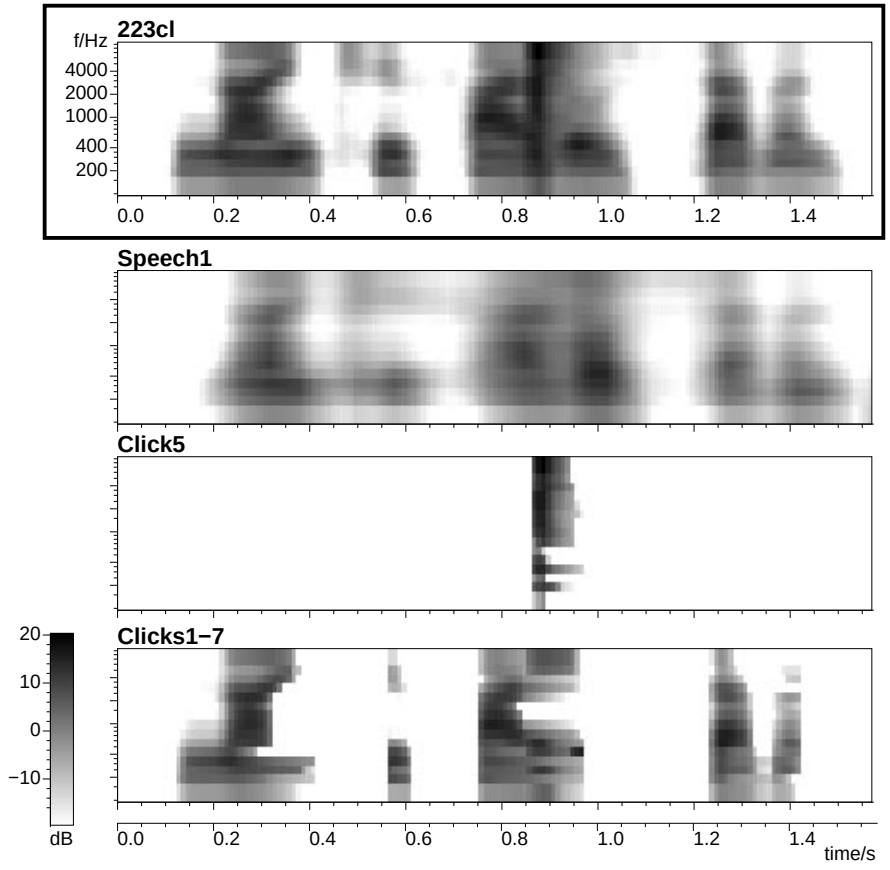
- **Reconstruct ‘canonical’ signal**



Example of speech & nonspeech



Putting it into the scene analyzer



- **Prediction shortfalls dominate residual**

5

Future work

- **Better signal predictions from the recognizer**
  - normalized training
  - weighted reconstruction
  - more classes?
- **Other immediate problems**
  - iteration!
  - starting hypothesis
  - granularity of integration
  - low-frequency separation
- **Nonspeech knowledge?**
- **Performance & evaluation**

Conclusions

- **Need to use scene analysis for real sounds**
- **Listeners' scene analysis relies on knowledge-based predictions**
- **Use prediction-driven formulation to employ speech-recognizer knowledge for explanation**
- **But: need better 'predictions'**
  - better inverse-classification
  - better normalization & inversion