Computational Auditory Scene Analysis exploiting Speech Recognition knowledge

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

Outline
1. Computational Auditory Scene Analysis
2. CASA for speech recognition
3. A speech hypothesis module
4. Speech & nonspeech examples
5. Current problems & future work

Current problems & future work

• Inaccurate reconstructions
  - predictions fail to account for all speech energy
• Iterating between speech/nonspeech
  - how best to use nonspeech estimates in ASR?
• Bootstrapping (start-up)
  - need to recognize speech in original mixture

Inaccurate reconstructions

• Problem:
  speech ‘prediction’ falls short of mixture energy
  → spurious nonspeech elements
• Solutions:
  1. More normalization → sharper models
     - spectral warping - multiscale normalization
     - ‘put back’ characteristics during reconstruction
  2. Less generalization → sharper models
     - more states e.g. context-dependent phones
     - less temporal smearing in features
  3. Condition on additional information
     - train NN with label class + ? input ? last state

Iterating between speech & nonspeech

• Central idea: iterative refinement of each component promotes separation
• Speech estimate guides nonspeech estimator by ‘predicting’ speech energy
• .. but how will speech recognizer be helped by good nonspeech estimates?
  - subtraction? does wrong thing; leaves holes
  - need new ‘masked’ acoustic score:
    p(X | q,Z)
    masking level from nonspeech
    (in spectral domain)

Bootstrapping

• Currently, first pass is speech recognizer
  - if speech is poorly recognized, will it converge?
  - unless speech is poorly recognized, why bother?
• Loss of pitch in ‘feature resynthesis’ is very prominent...
• How it should work:
  - recognizer trained on separated periodic/noise

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
  - better speech-hypothesis generation
Computational Auditory Scene Analysis exploiting Speech Recognition knowledge

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

Outline

1. Computational Auditory Scene Analysis
2. CASA for speech recognition
3. A speech hypothesis module
4. Speech & nonspeech examples
5. Current problems & future work
Current problems & future work

- **Inaccurate reconstructions**
  - predictions fail to account for all speech energy

- **Iterating between speech/nonspeech**
  - how best to use nonspeech estimates in ASR?

- **Bootstrapping (start-up)**
  - need to recognize speech in original mixture
Inaccurate reconstructions

- **Problem:**
  
  speech ‘prediction’ falls short of mixture energy
  → spurious nonspeech elements

- **Solutions:**
  
  1. **More normalization** → sharper models
     - spectral warping - multiscale normalization
     .. ‘put back’ characteristics during reconstruction
  
  2. **Less generalization** → sharper models
     - more states e.g. context-dependent phones
     - less temporal smearing in features
  
  3. **Condition on additional information**
     - train NN with label class + ? input ? last state

\[
E[\hat{X}_n | Q_n, f(X_n, \hat{X}_{n-1})]
\]
Iterating between speech & nonspeech

- Central idea: iterative refinement of each component promotes separation

- Speech estimate guides nonspeech estimator by ‘predicting’ speech energy

- .. but how will speech recognizer be helped by good nonspeech estimates?
  - subtraction? does wrong thing; leaves holes
  - need new ‘masked’ acoustic score:

\[ p(X \mid q,Z) \]

\textit{masking level from nonspeech}

(in spectral domain)
Bootstrapping

- Currently, first pass is speech recognizer
  - if speech is poorly recognized, will it converge?
  - unless speech is poorly recognized, why bother?

- Loss of *pitch* in ‘feature resynthesis’ is very prominent...

- How it should work:
  - recognizer trained on separated periodic/noise
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
  - better speech-hypothesis generation