Analysis-by-synthesis for source separation and speech recognition

Michael I Mandel
mim@mr-pc.org

Brooklyn College (CUNY)

Joint work with Young Suk Cho and Arun Narayanan (Ohio State)

Columbia Neural Network Seminar Series
September 8, 2015
1. Motivation: need for noise robustness
2. Non-parametric synthesis for speech enhancement
3. Parametric synthesis for speech recognition
4. Summary
Motivation: need for noise robustness

1. Need for better mobile voice quality
2. Need for noise robust automatic speech recognition (ASR)
3. Main challenge

Non-parametric synthesis for speech enhancement

Parametric synthesis for speech recognition

Summary
Motivation: need for noise robustness
- Need for better mobile voice quality
- Need for noise robust automatic speech recognition (ASR)
- Main challenge

Non-parametric synthesis for speech enhancement

Parametric synthesis for speech recognition

Summary
Motivation: need for noise robustness

Need for better mobile voice quality

- There are now more mobile devices than humans on earth\(^1\)
- But recording conditions for these devices leave much to be desired
- Can we recover high quality speech from noisy & degraded recordings?

\(^1\)http://www.independent.co.uk/life-style/gadgets-and-tech/news/there-are-officially-more-mobile-devices-than-people-in-the-world-9780518.html
Why mobile voice quality stinks

---

Motivation: need for noise robustness

Need for better mobile voice quality

Why mobile voice quality stinks

2 Jeff Hecht. Why mobile voice quality still stinks—and how to fix it. IEEE Spectrum, September 2014
Outline

1. Motivation: need for noise robustness
   - Need for better mobile voice quality
   - Need for noise robust automatic speech recognition (ASR)
   - Main challenge

2. Non-parametric synthesis for speech enhancement

3. Parametric synthesis for speech recognition

4. Summary
Conversational mobile software agents
Conversational mobile software agents need to work in

Source: Flickr user rickihuang
Motivation: need for noise robustness

Need for noise robust automatic speech recognition (ASR)

Conversational mobile software agents need to work in

Source: Flickr user retorta.net
Conversational mobile software agents need to work in
Motivation: need for noise robustness

Need for noise robust automatic speech recognition (ASR)

But automatic speech recognition doesn’t work there³

Motivation: need for noise robustness

- Need for better mobile voice quality
- Need for noise robust automatic speech recognition (ASR)
- Main challenge

Outline

1. Motivation: need for noise robustness
2. Non-parametric synthesis for speech enhancement
3. Parametric synthesis for speech recognition
4. Summary
Main challenge

Speech is a rich signal, it requires rich models
Speech is a rich signal, it requires rich models

- Synthesis models are rich enough to represent almost all speech
- Non-parametric synthesis models for high quality
  - DNN as non-linear distance function
- Parametric synthesis models for efficient representation
  - efficient gradient-based optimization of input (not model)
Outline

1 Motivation: need for noise robustness

2 Non-parametric synthesis for speech enhancement
   - Overview
   - Deep neural network as nonlinear distance function
   - Using this DNN for speech enhancement
   - Noise suppression experiments
   - Audio super-resolution experiments
   - Summary

3 Parametric synthesis for speech recognition

4 Summary
Outline

1 Motivation: need for noise robustness

2 Non-parametric synthesis for speech enhancement
   - Overview
   - Deep neural network as nonlinear distance function
   - Using this DNN for speech enhancement
   - Noise suppression experiments
   - Audio super-resolution experiments
   - Summary

3 Parametric synthesis for speech recognition

4 Summary
Concatenative resynthesis for speech enhancement\textsuperscript{4,5}

- Standard approaches try to modify noisy recordings
- We instead resynthesize a clean version of the same speech
- Should produce infinite suppression and high speech quality


Motivating example

- Your phone records your voice in quiet, close-talk conditions
- Uses those recordings to replace your voice in noisy, far-talk conditions
- Resynthesizes your speech from previous high-quality recordings
Concatenative resynthesis

- Use a large dictionary of \(~200\) ms “chunks” of audio
- Learn DNN-based affinity between dictionary & mixture chunks
- Perform concatenative synthesis of signal from dictionary
- General robust supervised nonlinear signal mapping framework

<table>
<thead>
<tr>
<th>Task</th>
<th>Map from</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise suppression</td>
<td>Noisy</td>
<td>Clean</td>
</tr>
<tr>
<td>Audio super-resolution</td>
<td>Reverberated, compressed</td>
<td>Clean</td>
</tr>
</tbody>
</table>
Outline

1 Motivation: need for noise robustness

2 Non-parametric synthesis for speech enhancement
   - Overview
   - Deep neural network as nonlinear distance function
     - Using this DNN for speech enhancement
     - Noise suppression experiments
     - Audio super-resolution experiments
     - Summary

3 Parametric synthesis for speech recognition

4 Summary
Deep neural network as nonlinear distance function

Generative

Mix

Clean

Data-intensive training
Hard to adapt

Discriminative

Mix

Mask

Moderate training data
Hard to adapt

Dictionary-based

Clean

Mix

Similarity

Data-efficient training
Very adaptable

---

Train DNN on correctly and incorrectly paired chunks

Noise

Clean

Neg

Pos

Mixture

Noise suppression
Train DNN on correctly and incorrectly paired chunks

Audio super-resolution
Outline

1. Motivation: need for noise robustness

2. Non-parametric synthesis for speech enhancement
   - Overview
   - Deep neural network as nonlinear distance function
   - Using this DNN for speech enhancement
   - Noise suppression experiments
   - Audio super-resolution experiments
   - Summary

3. Parametric synthesis for speech recognition

4. Summary
Find optimal sequence of clean chunks

\( \mathbf{x} = \{x_t\}_{t=0}^{T} \) input sequence of noisy chunks

\( \hat{\mathbf{z}} = \{z_t\}_{t=0}^{T} \) best sequence of corresponding dictionary chunks

\[
\hat{\mathbf{z}} = \arg\max_{\mathbf{z}} \prod_{t} p(z_t = j \mid x_t) \cdot p(z_t = j \mid z_{t-1} = i)
\]

\[
= \arg\max_{\mathbf{z}} \prod_{i} g(z_j, x_i) \cdot T_{ij}
\]
Find optimal sequence of clean chunks

- \( x = \{x_t\}_{t=0}^T \) input sequence of noisy chunks
- \( \hat{z} = \{z_t\}_{t=0}^T \) best sequence of corresponding dictionary chunks
- Affinity between clean and noisy chunks

\[
\hat{z} = \arg\max_z \prod_t p(z_t = j \mid x_t) \cdot p(z_t = j \mid z_{t-1} = i)
\]

\[
= \arg\max_z \prod_i g(z_j, x_i) \cdot T_{ij}
\]
Find optimal sequence of clean chunks

- $x = \{x_t\}_{t=0}^T$ input sequence of noisy chunks
- $\hat{z} = \{z_t\}_{t=0}^T$ best sequence of corresponding dictionary chunks

Affinity between clean and noisy chunks

Transition affinity between clean chunks

\[
\hat{z} = \arg\max_z \prod_t p(z_t = j \mid x_t) \ p(z_t = j \mid z_{t-1} = i)
\]

\[
= \arg\max_z \prod_i g(z_j, x_i) \ T_{ij}
\]
Compare all pairs of noisy and clean chunks
Compare all pairs of noisy and clean chunks
Compare all pairs of noisy and clean chunks
Non-parametric synthesis for speech enhancement

Using this DNN for speech enhancement

Compare all pairs of noisy and clean chunks

DNN

Observed mixture

D1
D2
D3
...

M4
M4
M4
...

Clean dictionary

Similarity

Michael Mandel (Brooklyn College)
Compare all pairs of noisy and clean chunks
Compare all pairs of noisy and clean chunks
Non-parametric synthesis for speech enhancement

Using this DNN for speech enhancement

Standard Viterbi algorithm for to find optimal sequence

Michael Mandel (Brooklyn College)
Standard Viterbi algorithm for to find optimal sequence
Outline

1 Motivation: need for noise robustness

2 Non-parametric synthesis for speech enhancement
   - Overview
   - Deep neural network as nonlinear distance function
   - Using this DNN for speech enhancement
   - Noise suppression experiments
   - Audio super-resolution experiments
   - Summary

3 Parametric synthesis for speech recognition

4 Summary
Original “clean” speech
Noisy speech
Traditional mask-based separation
Concatenative resynthesis output
Original “clean” speech
Subjective quality is high
Subjective quality is high

![Graph showing quality scores for different conditions (Clean, Concat, IRM NN, Noisy) across different quality metrics (Speech, Noise Sup, Overall). The graph indicates high subjective quality with scores above 90 in most cases.]
Subjective intelligibility is ok

![Graph showing words correctly identified vs. conditions (Clean, Concat, IRM NN, Noisy) for keywords and all words.](image-url)
Outline

1 Motivation: need for noise robustness

2 Non-parametric synthesis for speech enhancement
   - Overview
   - Deep neural network as nonlinear distance function
   - Using this DNN for speech enhancement
   - Noise suppression experiments
   - Audio super-resolution experiments
   - Summary

3 Parametric synthesis for speech recognition

4 Summary
Original clean speech
Reverberated, compressed, 20% packet loss
NMF-based bandwidth expansion output
Concatenative resynthesis output
Original clean speech
Subjective quality is high
Subjective quality is high

![Bar chart showing MUSHRA scores for different conditions]

- Clean
- Clean (hid)
- Rev
- Rev 8kHz
- RevOpusL20
- RevOpusL20 (hid)
- CleanAmr
- RevAmr
- RevOpusL20

Michael Mandel (Brooklyn College)
Subjective intelligibility is good
Outline

1. Motivation: need for noise robustness

2. Non-parametric synthesis for speech enhancement
   - Overview
   - Deep neural network as nonlinear distance function
   - Using this DNN for speech enhancement
   - Noise suppression experiments
   - Audio super-resolution experiments
   - Summary

3. Parametric synthesis for speech recognition

4. Summary
Summary

- Concatenative synthesizer, DNN as noise-robust selection function
- Instead of modifying noisy speech, replace it
  - completely eliminates noise, except for synthesis errors
  - produces high quality, natural-sounding speech
- General robust supervised nonlinear signal mapping framework
- Data-efficient to train and adaptable to new talkers
Future applications

- Generalize to audio-visual speech recognition
- Label dictionary elements ahead of time to enable
  - noise-robust non-parametric speech recognition
  - noise-robust pitch tracking
  - noise-robust speaker identification
- Incorporate language model into transition cost
- Develop efficient search mechanisms for large-vocabulary dictionaries
Outline

1. Motivation: need for noise robustness

2. Non-parametric synthesis for speech enhancement

3. Parametric synthesis for speech recognition
   - Overview
   - Algorithm
   - Results
   - Summary

4. Summary
Outline

1. Motivation: need for noise robustness
2. Non-parametric synthesis for speech enhancement
3. Parametric synthesis for speech recognition
   - Overview
   - Algorithm
   - Results
   - Summary
4. Summary
Mask-based source separation: Noisy
Mask-based source separation: Masked
Disrupts speech features: Noisy MFCCs

“He said such products would be marketed by other companies with experience him at this month.”
Disrupts speech features: Masked MFCCs

“He said such products would be marketed by other companies with experience him at this month.”
Disrupts speech features: Clean MFCCs

“He said such products would be marketed by other companies with experience in that business.”
Estimate better features using a strong prior model

“He said such products would be marketed by other companies with experience in that business.”
Our approach: Analysis-by-synthesis

- Synthesize speech signal so that it
  - looks like the observation
  - looks like speech
- Itakura-Saito divergence compares prediction with noisy observation
- Recognizer gives likelihood of speech-ness
- Both easy to optimize using gradient descent
Speech recognizer includes lots of information

Large vocabulary continuous speech recognizer captures:

- Acoustics of speech sounds
- The effect of neighboring speech sounds
- Pronunciation of words
- Order of words
Outline

1. Motivation: need for noise robustness
2. Non-parametric synthesis for speech enhancement
3. Parametric synthesis for speech recognition
   - Overview
   - Algorithm
   - Results
   - Summary
4. Summary
Optimization over speech features

- \( \mathbf{x} \): optimization state: MFCCs, \( \sim 10,000 \) dimensions
- \( y(\mathbf{x}) \): ASR features derived from \( \mathbf{x} \)
- \( M \): mask provided a priori by another source separator

\[
\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}; M) = \min_{\mathbf{x}} \left\{ (1 - \alpha) \mathcal{L}_I(\mathbf{x}; M) + \alpha \mathcal{L}_H(y(\mathbf{x})) \right\}
\]

- Total cost
Optimization over speech features

- \( \mathbf{x} \): optimization state: MFCCs, \( \sim 10,000 \) dimensions
- \( y(\mathbf{x}) \): ASR features derived from \( \mathbf{x} \)
- \( M \): mask provided a priori by another source separator

\[
\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}; M) = \min_{\mathbf{x}} \left\{ (1 - \alpha) \mathcal{L}_I(\mathbf{x}; M) + \alpha \mathcal{L}_H(y(\mathbf{x})) \right\}
\]

- Total cost
- Distance to noisy observation
Optimization over speech features

- $\mathbf{x}$: optimization state: MFCCs, $\sim 10,000$ dimensions
- $y(\mathbf{x})$: ASR features derived from $\mathbf{x}$
- $M$: mask provided a priori by another source separator

$$\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}; M) = \min_{\mathbf{x}} \left\{ (1 - \alpha) \mathcal{L}_I(\mathbf{x}; M) + \alpha \mathcal{L}_H(y(\mathbf{x})) \right\}$$

- Total cost
- Distance to noisy observation
- Negative log likelihood under recognizer
Analysis of audio meets resynthesis of MFCCs at mask
\( \mathcal{L}_1(x; M) \): Distance to noisy observation

- Resynthesize MFCCs to power spectrum, where mask was computed
- Do mask-aware comparison in that domain: weighted Itakura-Saito
  - between resynthesis, \( \tilde{S}_{\omega t}(x) \), and noisy observation, \( S \)
  - weighted by mask, \( M \)

\[
\mathcal{L}_1(x; M) = D_M(S \parallel \tilde{S}) = \sum_{\omega, t} M_{\omega t} \left( \frac{S_{\omega t}}{\tilde{S}_{\omega t}(x)} - \log \frac{S_{\omega t}}{\tilde{S}_{\omega t}(x)} - 1 \right)
\]

- Does not require modeling speech excitation
- Numerically differentiable with respect to \( x \)
$\mathcal{L}_H(y(x))$: Likelihood under recognizer

- Large vocabulary continuous speech recognizer
  - big hidden Markov model (HMM)
  - approximated by the lattice of likely paths
- Closed form gradient with respect to $x$
- Serves as a model of clean MFCC sequences
\( \mathcal{L}_H(y(x)) \): Likelihood under recognizer

- Large vocabulary continuous speech recognizer
  - big hidden Markov model (HMM)
  - approximated by the lattice of likely paths
- Closed form gradient with respect to \( x \)
- Serves as a model of clean MFCC sequences
Optimization

- State space of approximately $13 \times 800 \approx 10,000$ dimensions
- Quasi-Newton optimization, BFGS
  - gradient plus approximate second-order information
- Closed form gradient of HMM likelihood
  - using a forward-backward algorithm
- Numerical gradient of IS divergence
  - independent costs and gradients for each frame
Outline

1. Motivation: need for noise robustness
2. Non-parametric synthesis for speech enhancement
3. Parametric synthesis for speech recognition
   - Overview
   - Algorithm
   - Results
   - Summary
4. Summary
Experiment

- AURORA4 corpus
  - read Wall Street Journal sentences (5000 word vocabulary)
  - six environmental noise types
  - SNRs between 5 and 15 dB
- Masks from ideal binary mask and estimated ratio mask\(^7\)

---

### Recognition results

- Word error rate (%) averaged across noise type

<table>
<thead>
<tr>
<th></th>
<th>Mask</th>
<th>Direct</th>
<th>A-by-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>30.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated</td>
<td>16.18</td>
<td>15.31</td>
<td></td>
</tr>
<tr>
<td>Oracle</td>
<td>14.38</td>
<td>13.62</td>
<td></td>
</tr>
<tr>
<td>Clean</td>
<td>9.54</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reconstruction results

- Itakura-Saito divergence between resynthesized speech and original

<table>
<thead>
<tr>
<th>Mask</th>
<th>Direct</th>
<th>A-by-S</th>
<th>$\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>272301</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated</td>
<td>276497</td>
<td>275224</td>
<td>-1273</td>
</tr>
<tr>
<td>Oracle</td>
<td>273006</td>
<td>272506</td>
<td>-500</td>
</tr>
</tbody>
</table>
Resynthesis gets closer to reliable regions
Resynthesis gets closer to reliable regions
Resynthesis gets closer to reliable regions
Resynthesis gets closer to reliable regions
Outline

1. Motivation: need for noise robustness

2. Non-parametric synthesis for speech enhancement

3. Parametric synthesis for speech recognition
   - Overview
   - Algorithm
   - Results
   - Summary

4. Summary
Summary

- Use a full recognizer as a prior model for clean speech
- Synthesize from MFCCs to the domain of the mask
- Adjust synthesis of speech signal so that it
  - looks like the observation
  - looks like speech
- Reduces recognition errors, distance to clean utterance
Future directions

- Apply to DNN-based acoustic models
- Model speech excitation for full resynthesis of clean speech
- Model multiple simultaneous speakers and estimate masks jointly
- Combine with similar binaural model to include spatial clustering
Outline

1. Motivation: need for noise robustness
2. Non-parametric synthesis for speech enhancement
3. Parametric synthesis for speech recognition
4. Summary
Summary

- Synthesizers provide strong prior information
- Non-parametric synthesis models for high quality
  - learned nonlinear matching function for perceptually motivated features
- Parametric synthesis models for efficient representation
  - strong, differentiable prior model of speech
Synthesizers provide strong prior information

- Non-parametric synthesis models for high quality
  - learned nonlinear matching function for perceptually motivated features

- Parametric synthesis models for efficient representation
  - strong, differentiable prior model of speech

Thanks!
Summary

- Synthesizers provide strong prior information
- Non-parametric synthesis models for high quality
  - learned nonlinear matching function for perceptually motivated features
- Parametric synthesis models for efficient representation
  - strong, differentiable prior model of speech

Thanks!
Any questions?
Outline

5 Parametric synthesis for separation
Re-estimate mask using resynthesis: Original
Re-estimate mask using resynthesis: Re-estimate