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

Outline

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

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

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

Need for better mobile voice quality

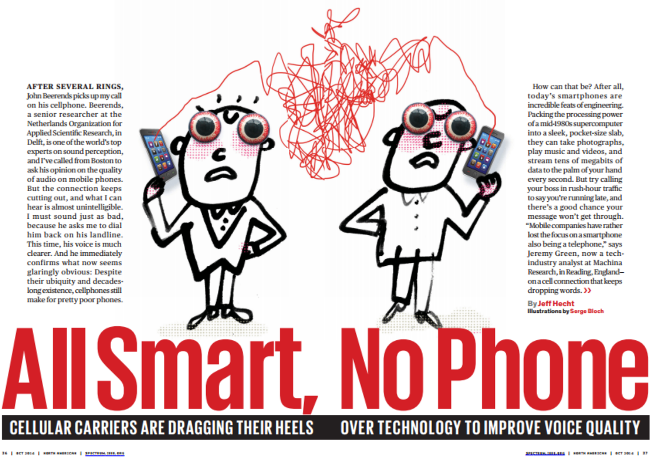
- There are now more mobile devices than humans on earth[1]
- But recording conditions for these devices leave much to be desired
- Can we recover high quality speech from noisy & degraded recordings?

[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[2]

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

Why mobile voice quality stinks[2]



AFTER SEVERAL RINGS, John Beerends picks up my call on his cellphone. Beerends, a senior researcher at the Netherlands Organization for Applied Scientific Research, in Delft, is one of the world's top experts on sound perception, and I've called from Boston to ask his opinion on the quality of audio on mobile phones. But the connection keeps cutting out, and what I can hear is almost unintelligible. I must sound just as bad, because he asks me to dial him back on his landline. This time, his voice is much clearer. And he immediately confirms what now seems glaringly obvious: Despite their ubiquity and decades-long existence, cellphones still make for pretty poor phones.

How can that be? After all, today's smartphones are incredible feats of engineering. Packing the processing power of a mid-1980s supercomputer into a sleek, pocket-size slab, they can take photographs, play music and videos, and stream tens of megabits of data to the palm of your hand every second. But try calling your boss in rush-hour traffic to say you're running late, and there's a good chance your message won't get through. "Mobile companies have rather lost the focus on a smartphone also being a telephone," says Jeremy Green, now a tech-industry analyst at Machina Research, in Reading, England–on a cell connection that keeps dropping words. »

By Jeff Hecht
Illustrations by Serge Bloch

All Smart, No Phone

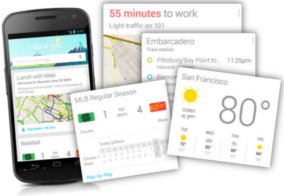
CELLULAR CARRIERS ARE DRAGGING THEIR HEELS OVER TECHNOLOGY TO IMPROVE VOICE QUALITY

[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



Source: Tom Vanleenhove

Conversational mobile software agents need to work in

Source: Flickr user rickihuang

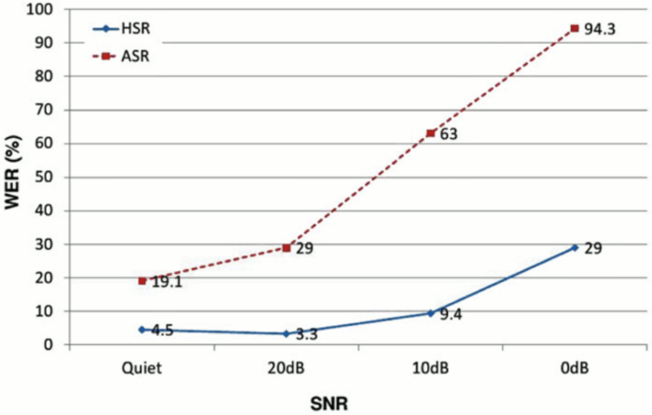
Conversational mobile software agents need to work in

Source: Flickr user retorta_net

Conversational mobile software agents need to work in

Source: Flickr user Brian_Indrelunas

But automatic speech recognition doesn't work there[3]



(b) 8000 word vocabulary

[3] Amit Juneja. A comparison of automatic and human speech recognition in null grammar. *The Journal of the Acoustical Society of America*, 131(3):EL256–EL261, February 2012

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

Main challenge

Speech is a rich signal, it requires rich models

Main challenge

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

Outline

Concatenative resynthesis for speech enhancement[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

[4] Michael I Mandel, Young-Suk Cho, and Yuxuan Wang. Learning a concatenative resynthesis system for noise suppression. In *Proc. IEEE GlobalSIP*, 2014

[5] Michael I Mandel and Young Suk Cho. Audio super-resolution using concatenative resynthesis. In *Proc. IEEE WASPAA*, 2015. To appear

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

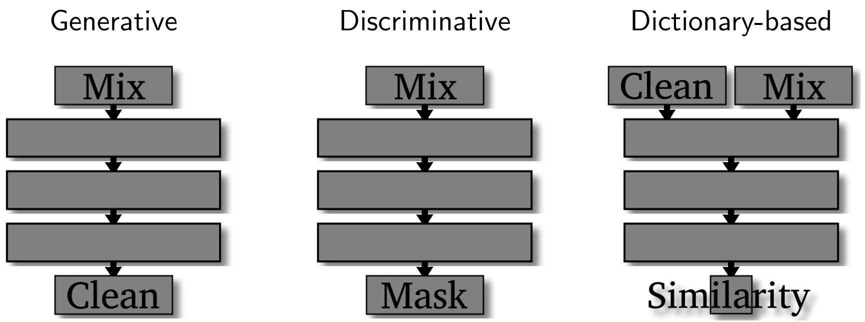
Concatenative resynthesis

- 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

Task	Map from	To
Noise suppression	Noisy	Clean
Audio super-resolution	Reverberated, compressed	Clean

Outline

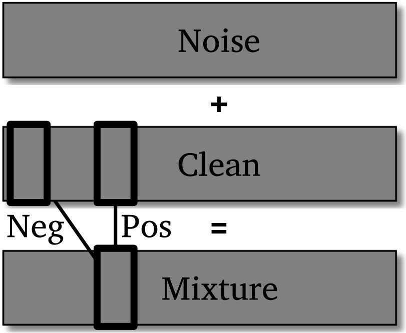
Deep neural network as nonlinear distance function[6]



Generative	Discriminative	Dictionary-based
Data-intensive training	Moderate training data	Data-efficient training
Hard to adapt	Hard to adapt	Very adaptable

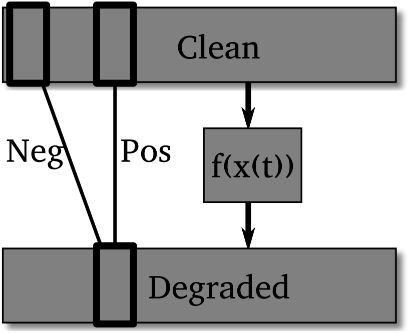
[6] Michael I Mandel, Young-Suk Cho, and Yuxuan Wang. Learning a concatenative resynthesis system for noise suppression. In *Proc. IEEE GlobalSIP*, 2014

Train DNN on correctly and incorrectly paired chunks



Noise suppression

Train DNN on correctly and incorrectly paired chunks



Audio super-resolution

Outline

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

$$\begin{aligned}\hat{\mathbf{z}} &= \operatorname*{argmax}_{\mathbf{z}} \prod_t p(z_t = j \mid x_t) \;\; p(z_t = j \mid z_{t-1} = i) \\ &= \operatorname*{argmax}_{\mathbf{z}} \prod_i g(z_j, x_i) \;\; T_{ij}\end{aligned}$$

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

- Affinity between clean and noisy chunks

$$\begin{aligned}\hat{\mathbf{z}} &= \operatorname*{argmax}_{\mathbf{z}} \prod_t \; p(z_t = j \mid x_t) \;\; p(z_t = j \mid z_{t-1} = i) \\ &= \operatorname*{argmax}_{\mathbf{z}} \prod_i \; g(z_j, x_i) \;\; T_{ij}\end{aligned}$$

Find optimal sequence of clean chunks

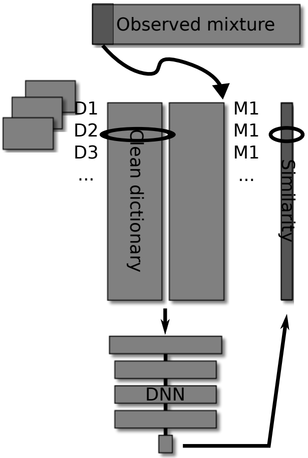
- $\mathbf{x} = \{x_t\}_{t=0}^{T}$ input sequence of noisy chunks
- $\hat{\mathbf{z}} = \{z_t\}_{t=0}^{T}$ best sequence of corresponding dictionary chunks

- Affinity between clean and noisy chunks
- Transition affinity between clean chunks

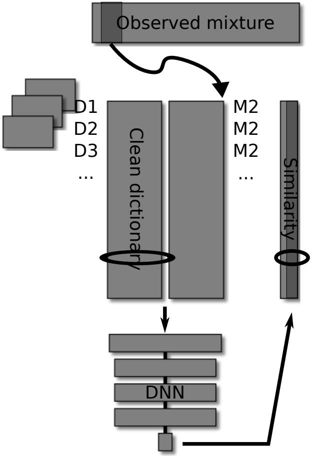
$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}} \prod_t \; p(z_t = j \mid x_t) \;\; p(z_t = j \mid z_{t-1} = i)$$

$$= \underset{\mathbf{z}}{\operatorname{argmax}} \prod_i \; g(z_j, x_i) \;\; T_{ij}$$

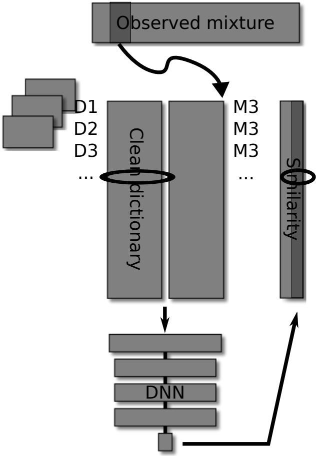
Compare all pairs of noisy and clean chunks



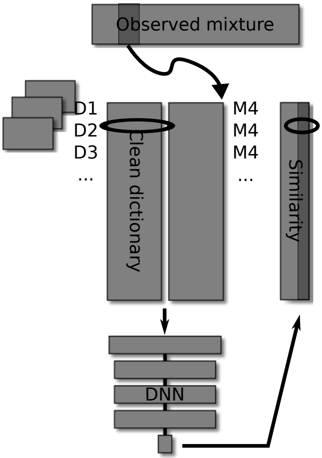
Compare all pairs of noisy and clean chunks



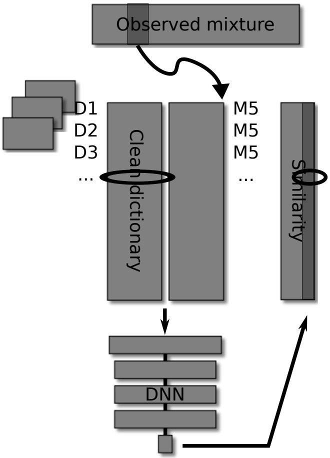
Compare all pairs of noisy and clean chunks



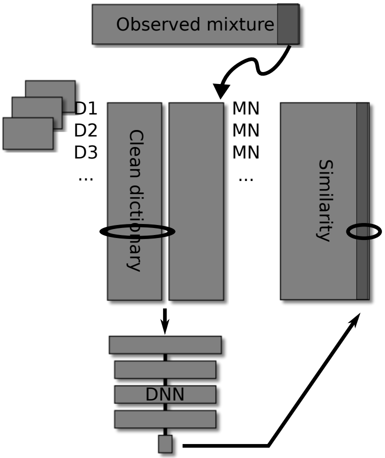
Compare all pairs of noisy and clean chunks



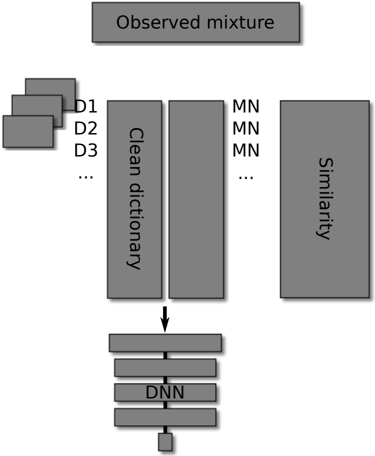
Compare all pairs of noisy and clean chunks



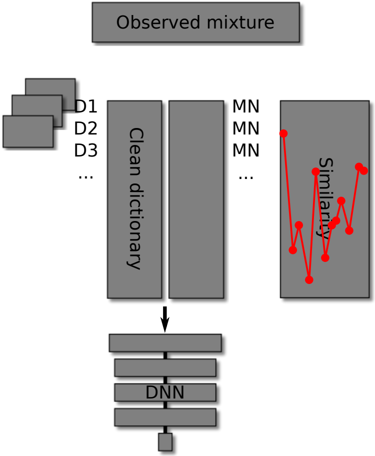
Compare all pairs of noisy and clean chunks



Standard Viterbi algorithm for to find optimal sequence

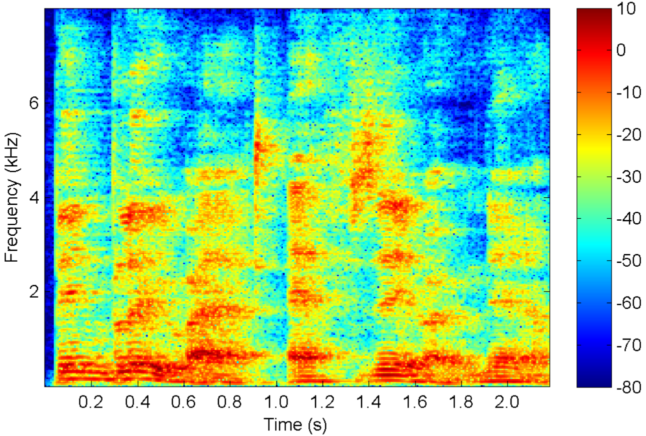


Standard Viterbi algorithm for to find optimal sequence

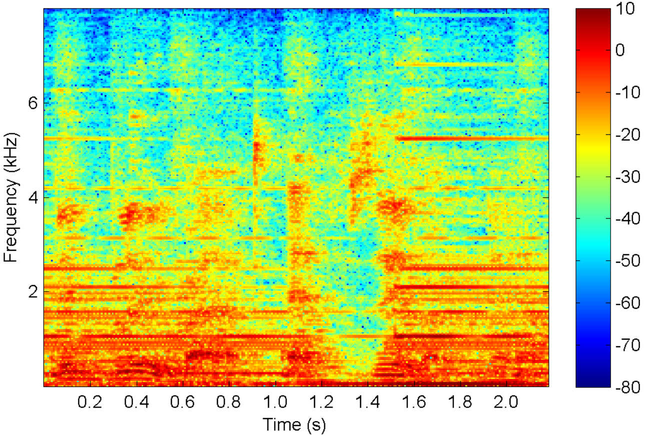


Outline

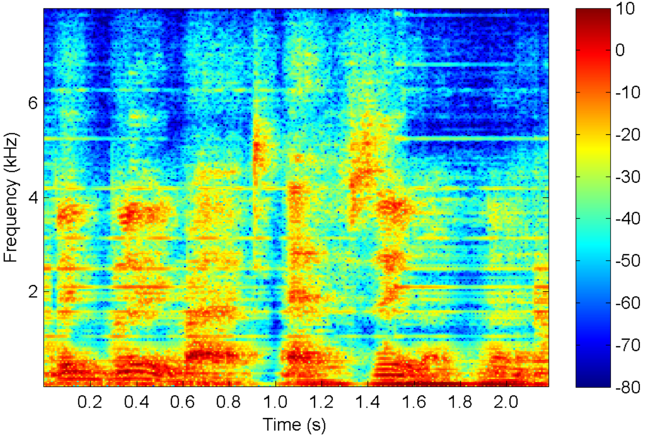
Original "clean" speech



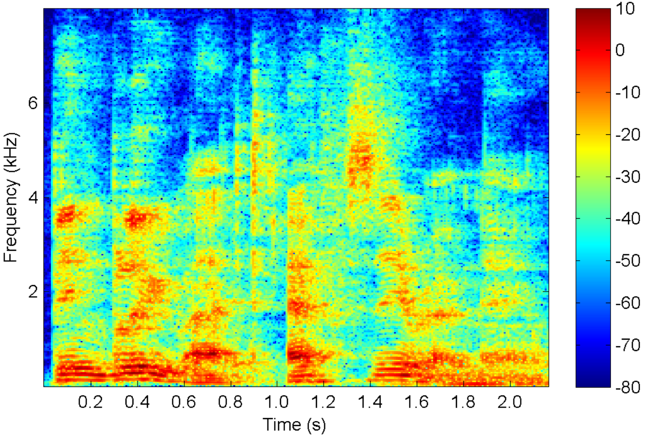
Noisy speech



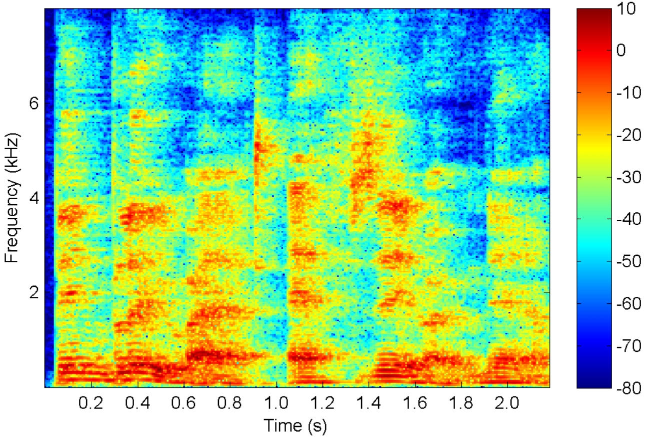
Traditional mask-based separation



Concatenative resynthesis output

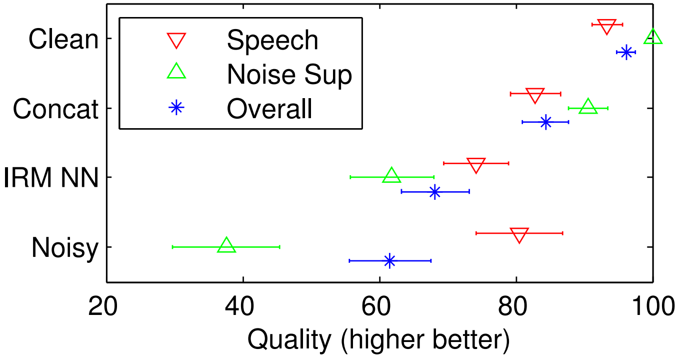


Original "clean" speech

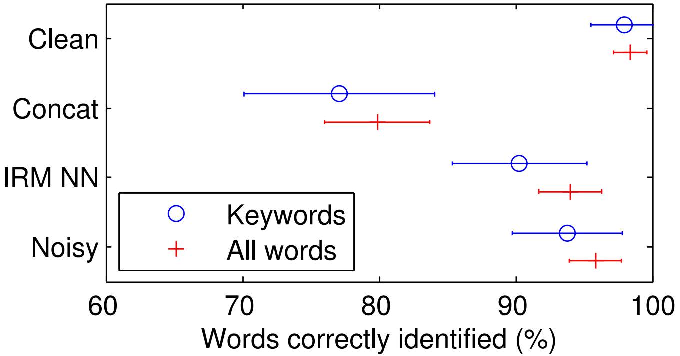


Subjective quality is high

Subjective quality is high



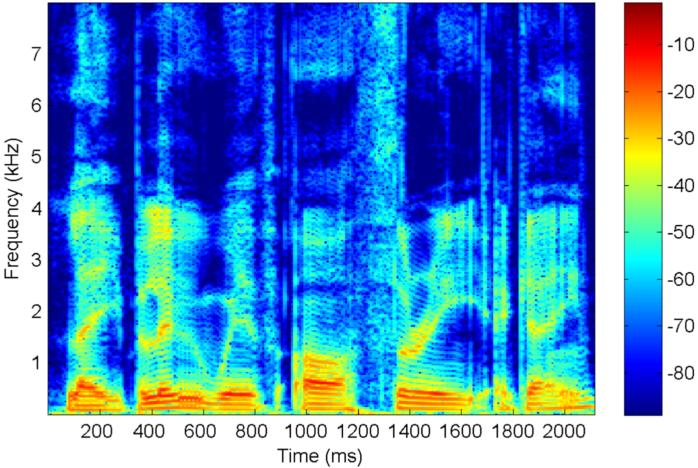
Subjective intelligibility is ok



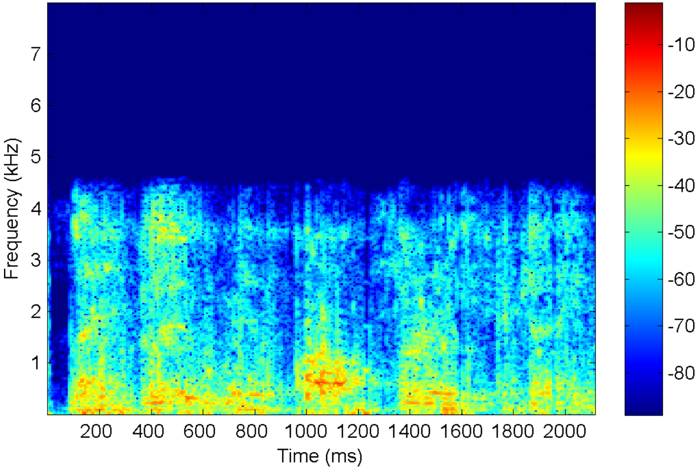
Outline

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

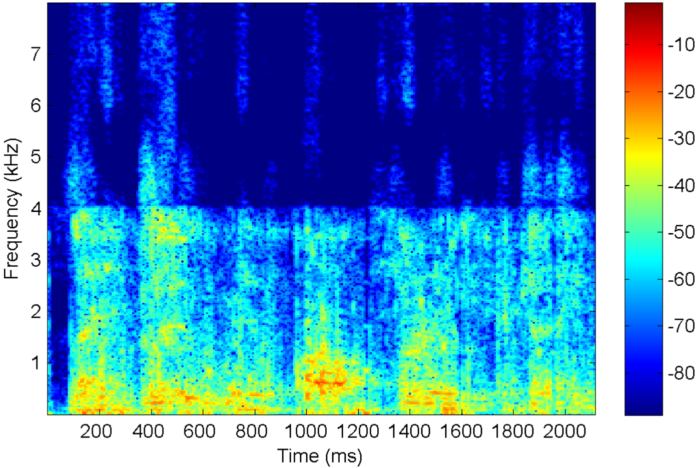
Original clean speech



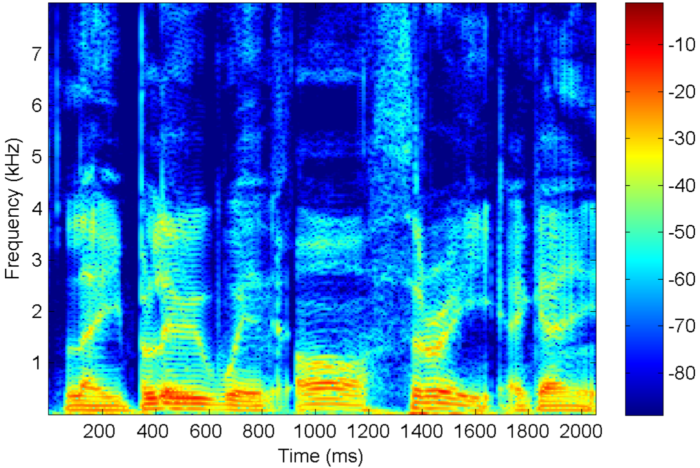
Reverberated, compressed, 20% packet loss



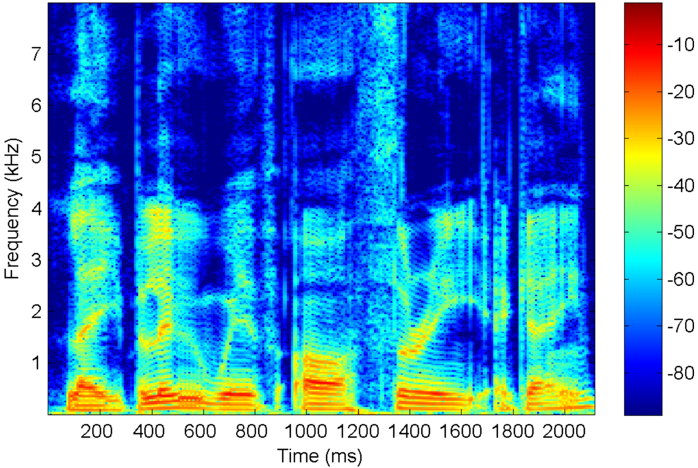
NMF-based bandwidth expansion output



Concatenative resynthesis output

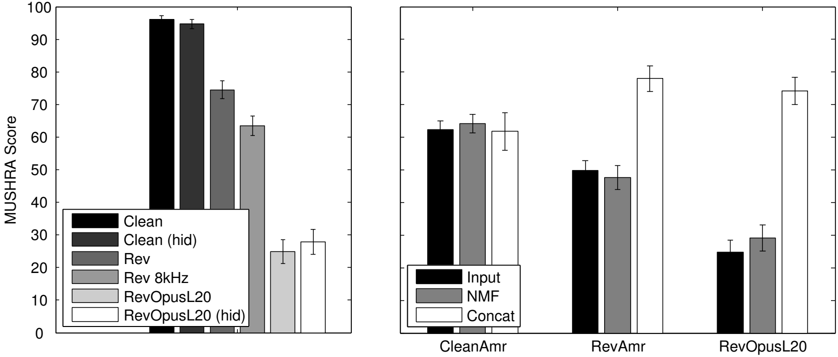


Original clean speech

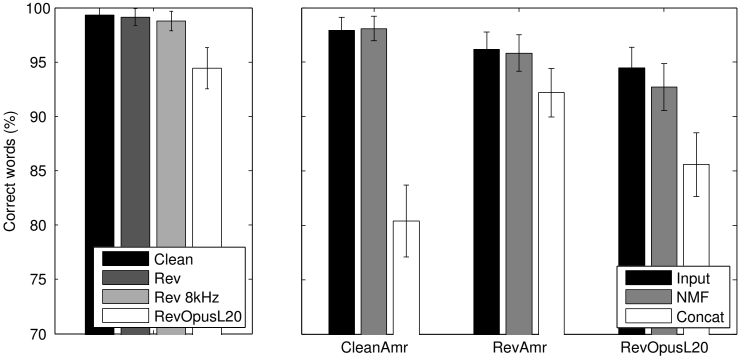


Subjective quality is high

Subjective quality is high



Subjective intelligibility is good



Outline

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

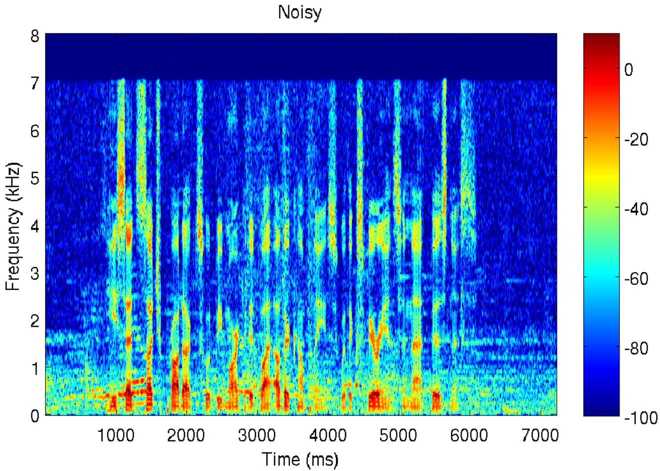
- Generalize to audio-visual speech recognition
- Label dictionary elements ahead of time to enable
  - noise-robust non-parametric speech recognition
  - noise-robust pitch tracking
  - noise-robust speaker identification
- Incorporate language model into transition cost
- Develop efficient search mechanisms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

Outline

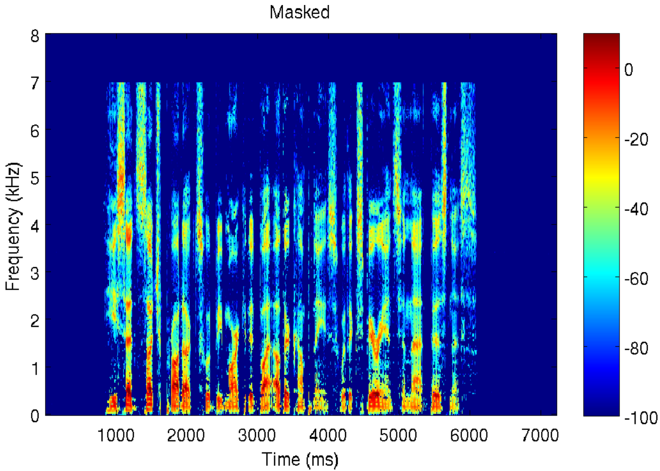
Outline

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

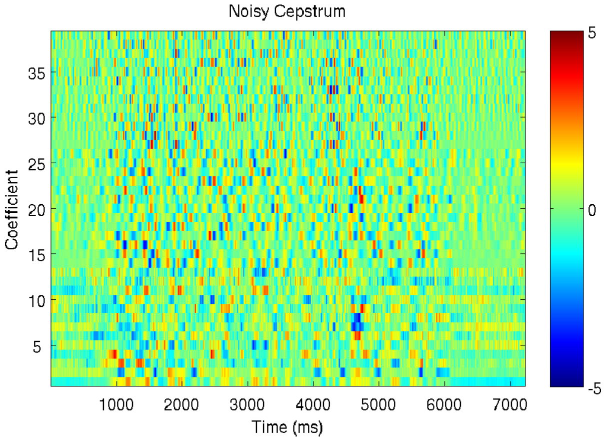
Mask-based source separation: Noisy



Mask-based source separation: Masked

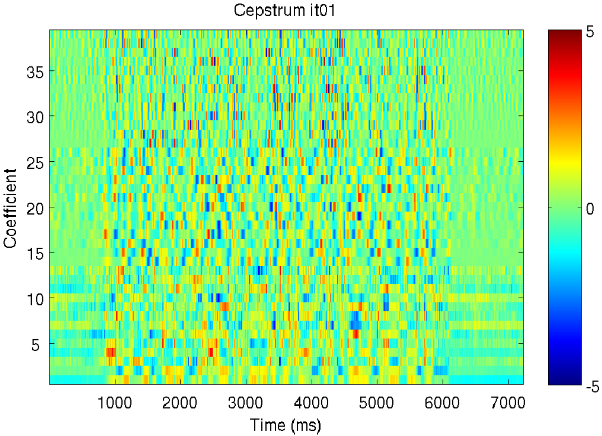


Disrupts speech features: Noisy MFCCs



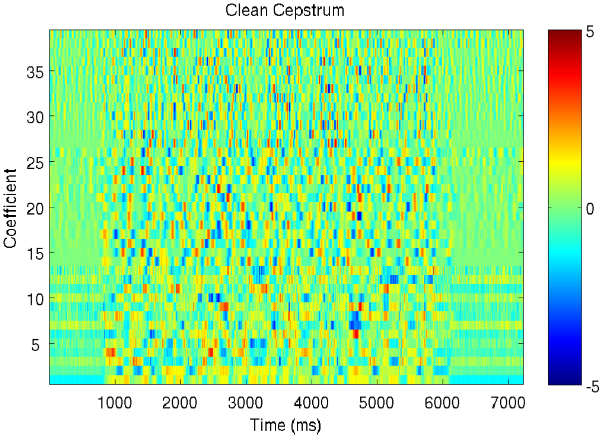
"He said such products would be marketed by other companies with experience ~~him at this month~~."

Disrupts speech features: Masked MFCCs



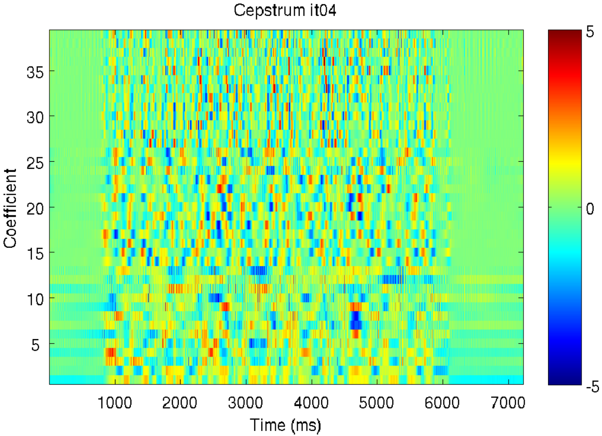
"He said such products would be marketed by other companies with experience ~~him at this month~~."

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

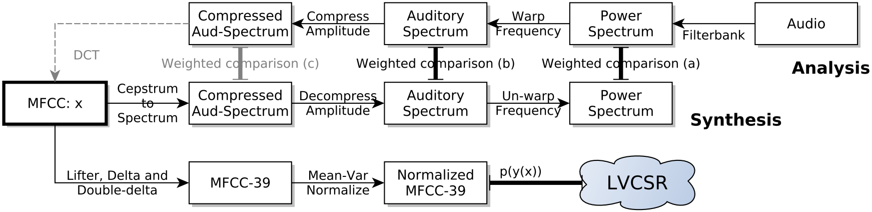
Optimization over speech features

- $\mathbf{x}$: 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



$\mathcal{L}_I(\mathbf{x}; M)$: Distance to noisy observation

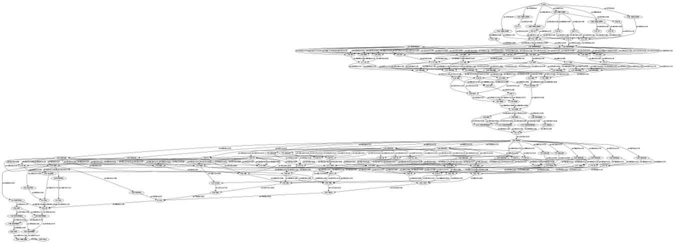
- Resynthesize MFCCs to power spectrum, where mask was computed
- Do mask-aware comparison in that domain: weighted Itakura-Saito
  - between resynthesis, $\tilde{S}_{\omega t}(\mathbf{x})$, and noisy observation, S
  - weighted by mask, M

$$\mathcal{L}_I(\mathbf{x}; M) = D_M(S \,\|\, \tilde{S}) = \sum_{\omega,t} M_{\omega t} \left(\frac{S_{\omega t}}{\tilde{S}_{\omega t}(\mathbf{x})} - \log \frac{S_{\omega t}}{\tilde{S}_{\omega t}(\mathbf{x})} - 1 \right)$$

- Does not require modeling speech excitation
- Numerically differentiable with respect to $\mathbf{x}$

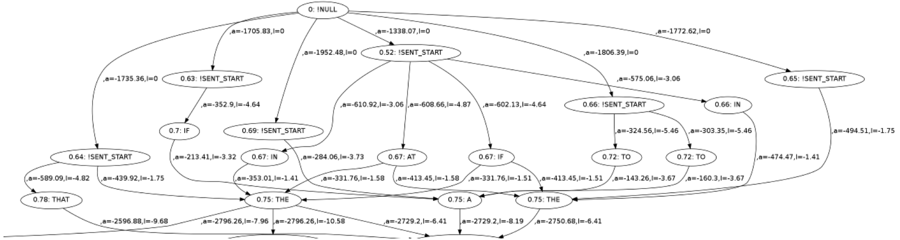
$\mathcal{L}_H(y(\mathbf{x}))$: Likelihood under recognizer

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



$\mathcal{L}_H(y(\mathbf{x}))$: 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 $\mathbf{x}$
- 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

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]

[7] Arun Narayanan and DeLiang Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7092–7096. IEEE, May 2013

Recognition results

- Word error rate (%) averaged across noise type

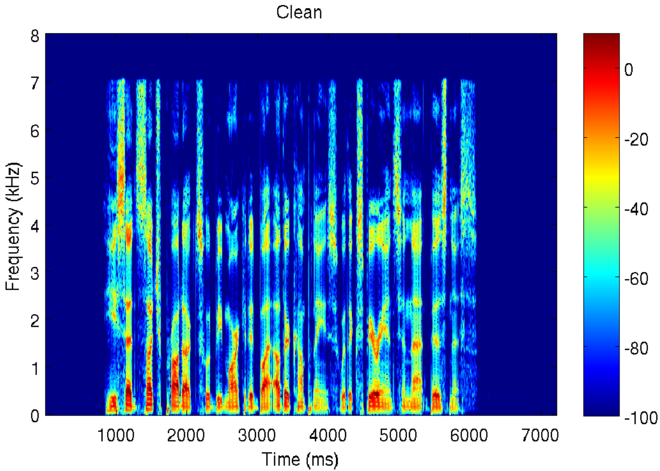
Mask	Direct	A-by-S
Noisy	30.94	
Estimated	16.18	15.31
Oracle	14.38	13.62
Clean	9.54	

Reconstruction results

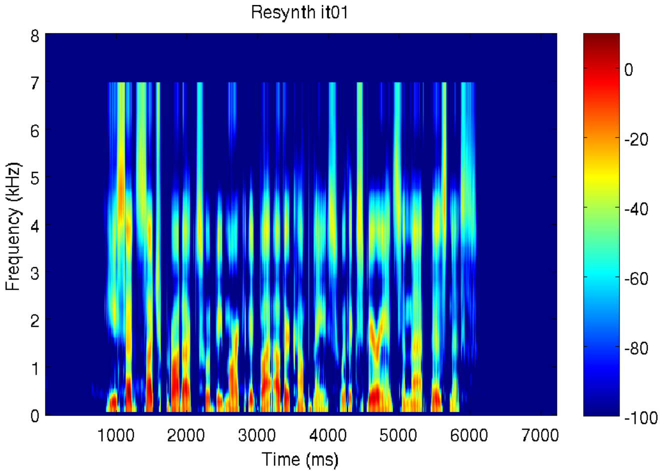
- Itakura-Saito divergence between resynthesized speech and original

Mask	Direct	A-by-S	Δ
Noisy	272301		
Estimated	276497	275224	-1273
Oracle	273006	272506	-500

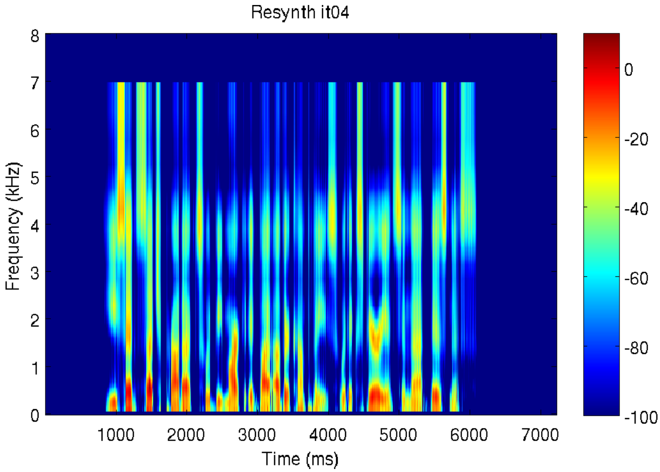
Resynthesis gets closer to reliable regions



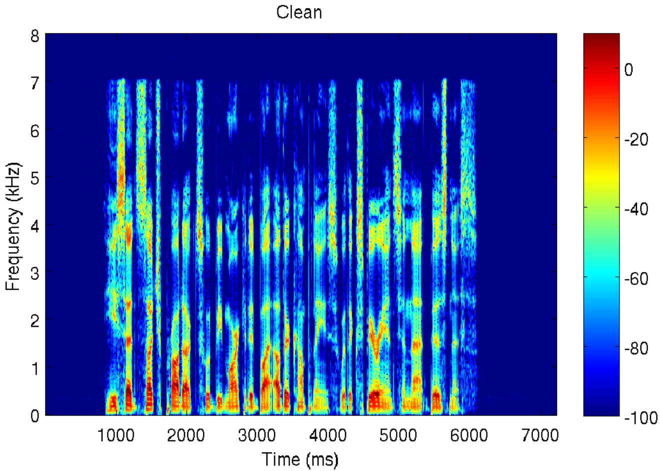
Resynthesis gets closer to reliable regions



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

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

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!

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

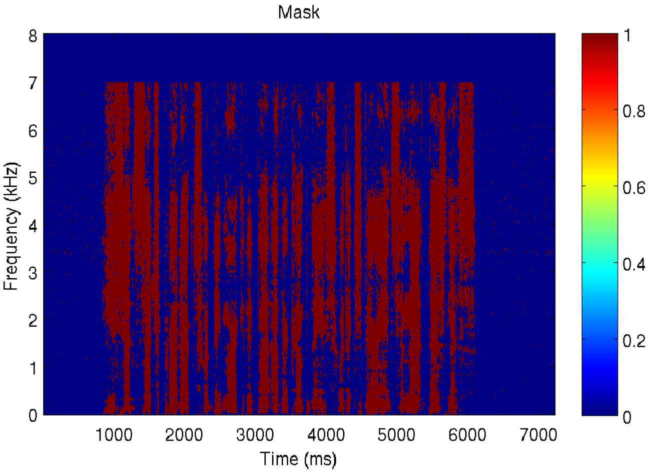
Thanks!

Any questions?

Outline

5. Parametric synthesis for separation

Re-estimate mask using resynthesis: Original



Re-estimate mask using resynthesis: Re-estimate