# Audio super-resolution using concatenative resynthesis

Michael I Mandel & Young Suk Cho
Brooklyn College (CUNY) & Ohio State

mim@mr-pc.org

WASPAA
October 20, 2015

# Audio super-resolution

- Imagine you are talking to someone on the phone

# Audio super-resolution

- Imagine you are talking to someone on the phone
- They are using hands free mode

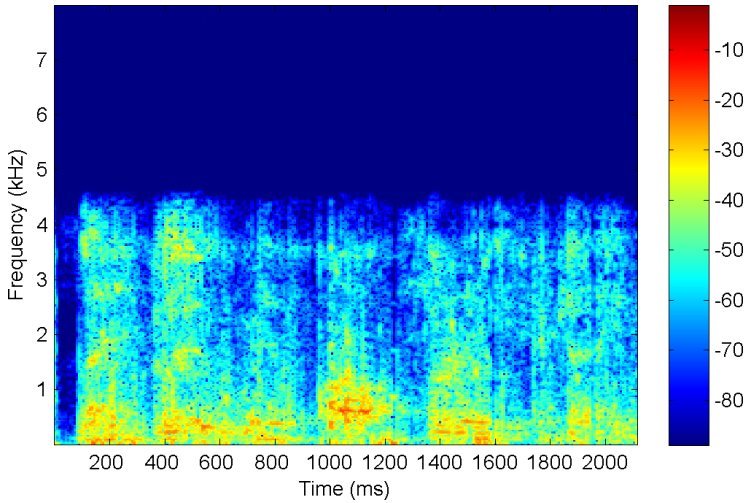
# Audio super-resolution

- Imagine you are talking to someone on the phone
- They are using hands free mode
- You have bad reception

# Audio super-resolution

- Imagine you are talking to someone on the phone
- They are using hands free mode
- You have bad reception

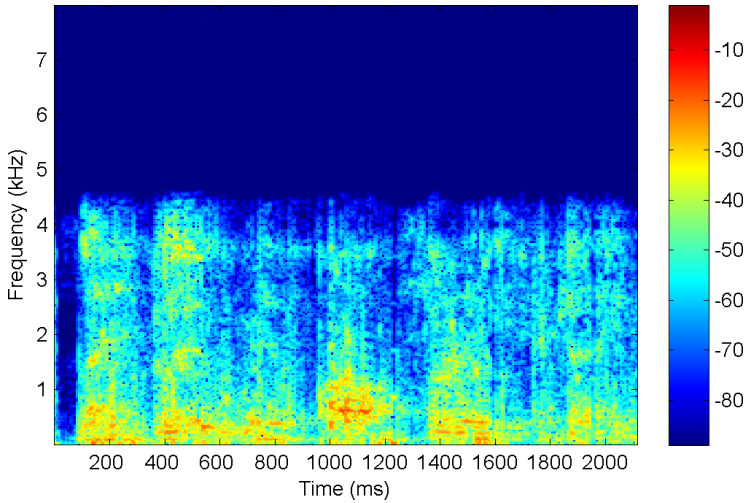
Have



# Audio super-resolution

- Imagine you are talking to someone on the phone
- They are using hands free mode
- You have bad reception
- Can we make it sound clean, wideband, and close-talk?

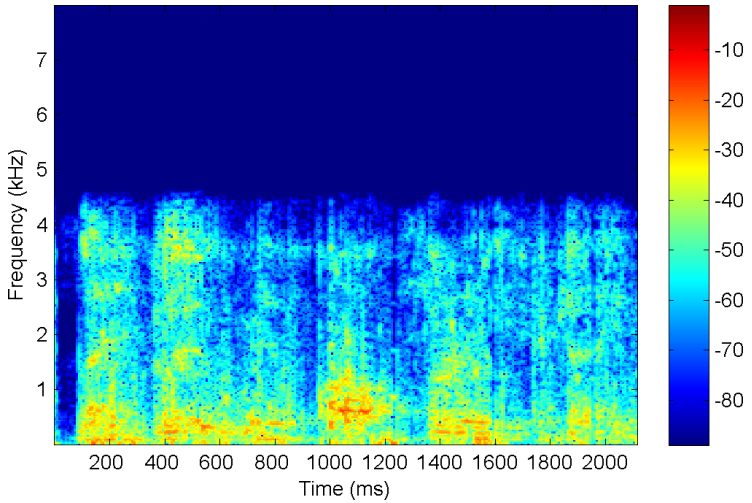
Have



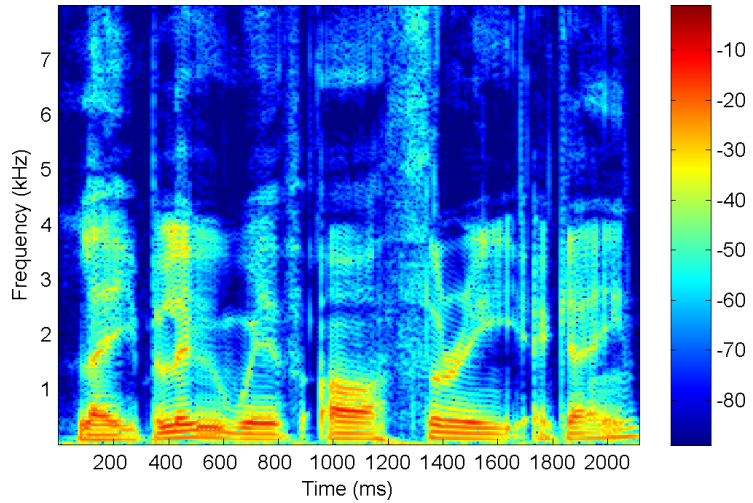
# Audio super-resolution

- Imagine you are talking to someone on the phone
- They are using hands free mode
- You have bad reception
- Can we make it sound clean, wideband, and close-talk?

Have




Want



# Concatenative resynthesis

- Standard approaches try to modify bad recordings
- We instead resynthesize a clean version of the same speech
  - synthesis systems model speech well, and only speech
  - should produce infinite suppression and high speech quality
- Previously introduced for noise suppression[1]

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

# Concatenative resynthesis is a general nonlinear mapping

- Perform concatenative synthesis of signal from dictionary
  - use a large dictionary of ∼200 ms "chunks" of audio
  - learn DNN-based affinity between dictionary & mixture chunks
- General robust supervised nonlinear signal mapping framework

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

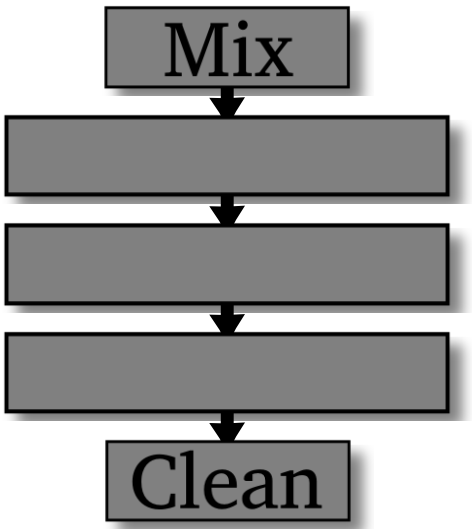
# Outline

1. Deep neural network as nonlinear selection function
2. Using this DNN for speech enhancement
3. Audio super-resolution experiments
4. Summary

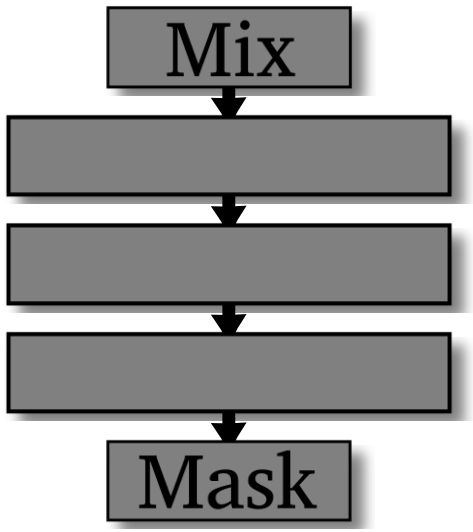
# Outline

# Deep neural network as nonlinear selection function

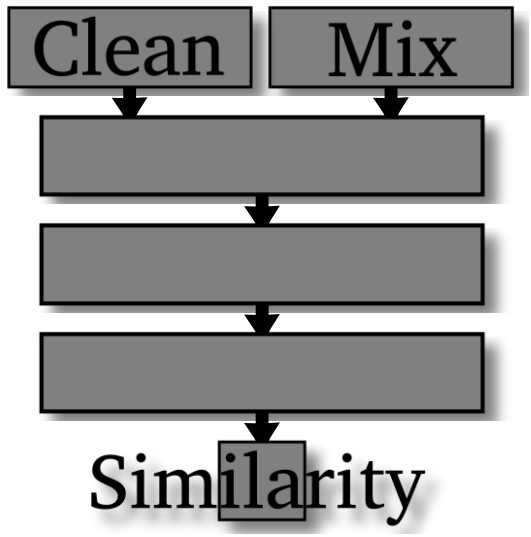
**Regression**
Data-intensive training
New talker hard



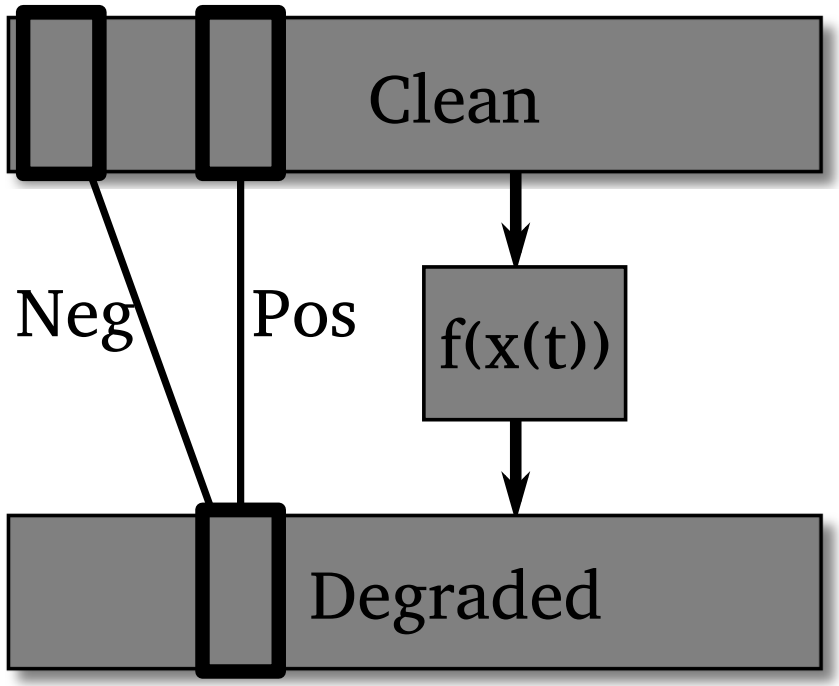

**Classification**
Moderate training data
New talker hard




**Selection**
Data-efficient training
New talker easy



# Train DNN on correctly and incorrectly paired chunks



# 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

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

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

# 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

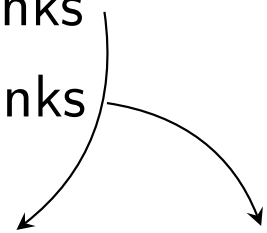
$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}} \prod_{t} p(z_t = j \mid x_t) \; p(z_t = j \mid z_{t-1} = i)$$

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

# 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

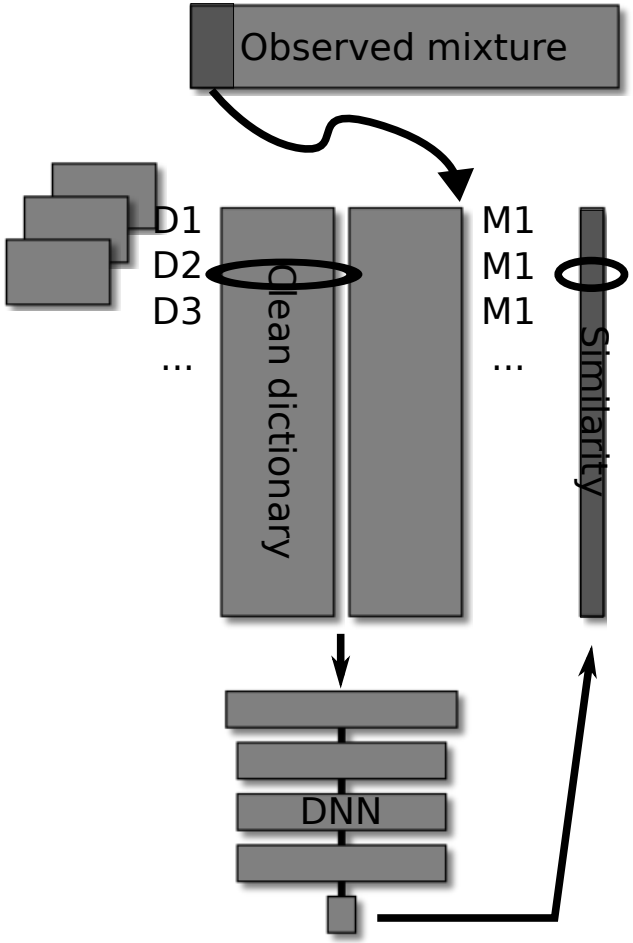
- Affinity between clean and noisy chunks
- Transition affinity between clean chunks



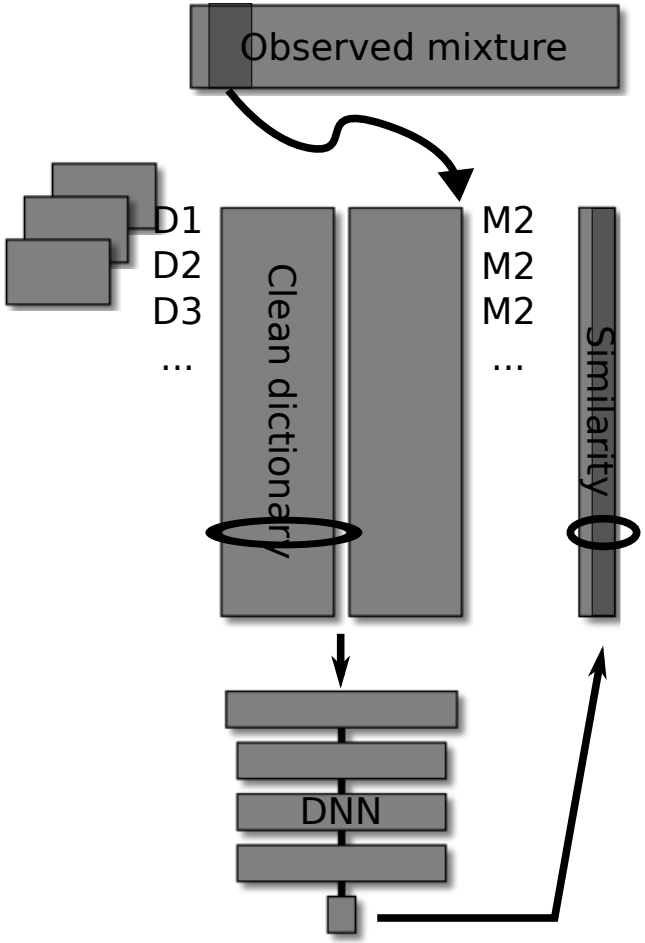
$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}} \prod_{t} p(z_t = j \,|\, x_t) \; p(z_t = j \,|\, z_{t-1} = i)$$

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

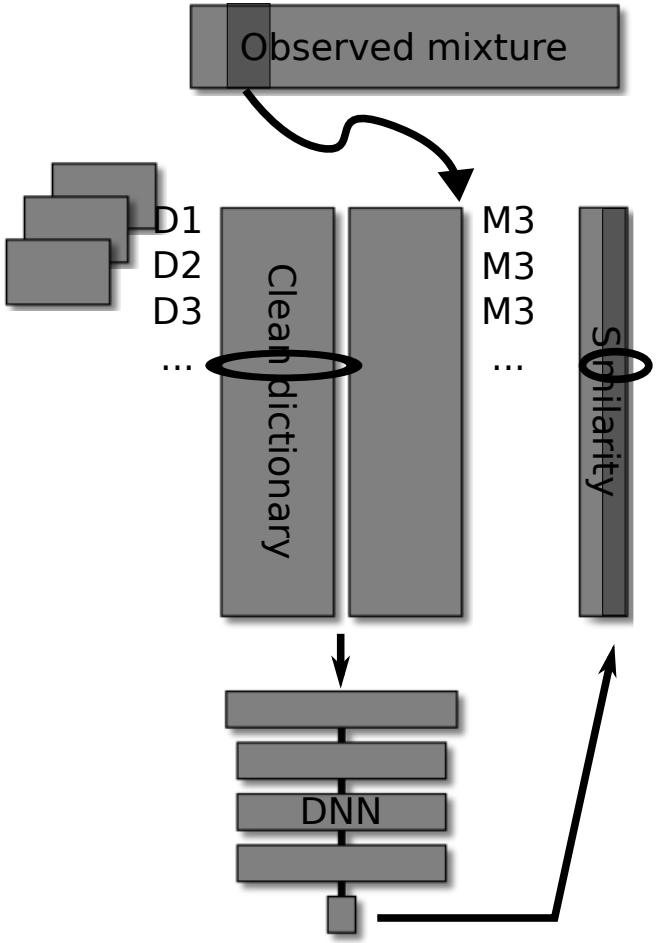
# Compare all pairs of noisy and clean chunks



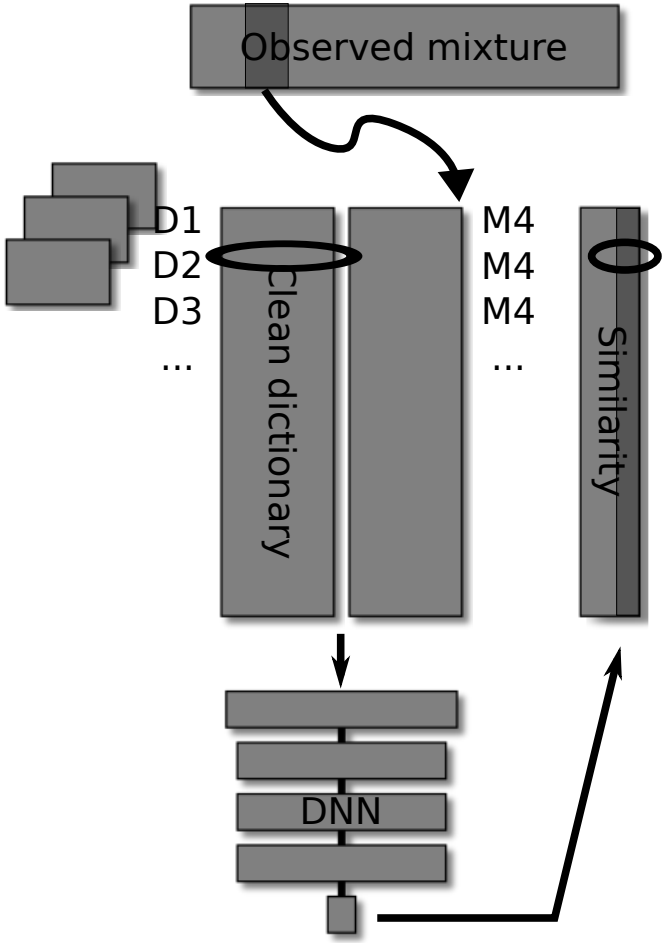
# Compare all pairs of noisy and clean chunks



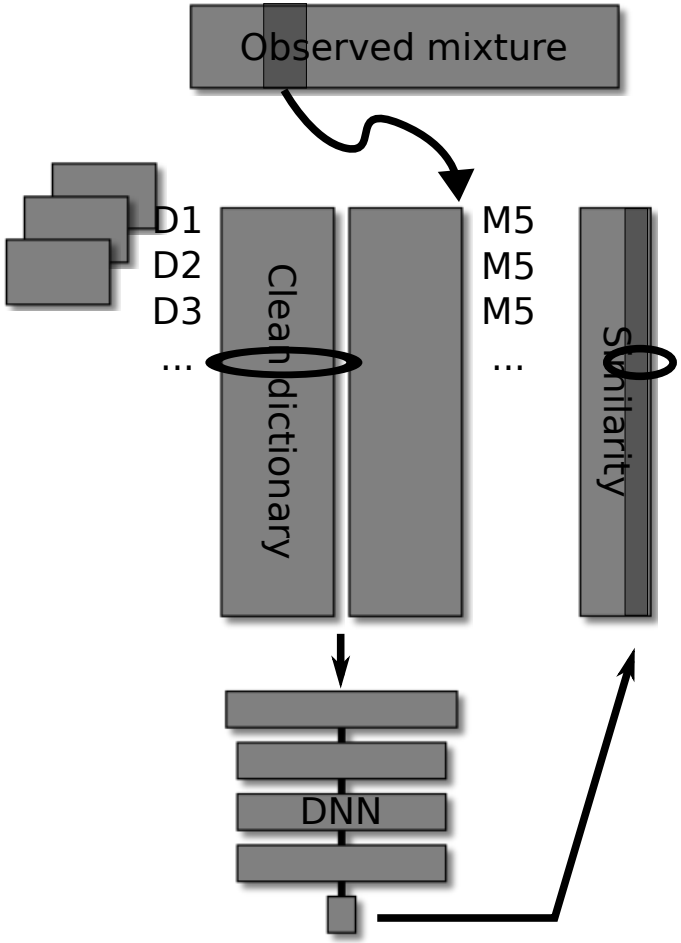
# Compare all pairs of noisy and clean chunks



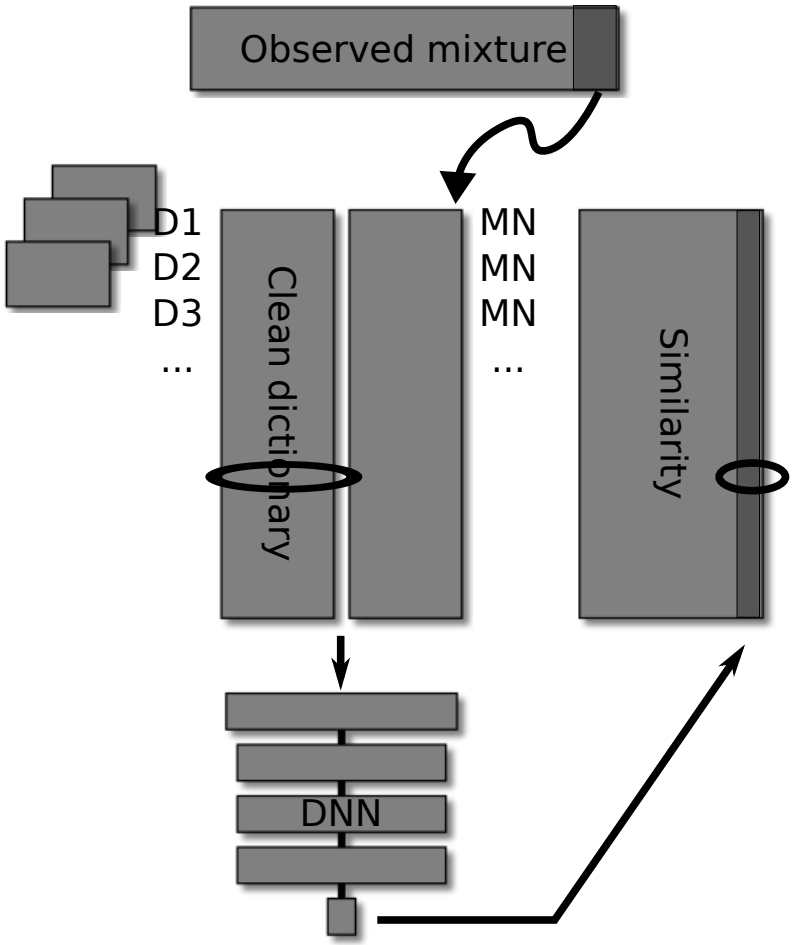
# Compare all pairs of noisy and clean chunks



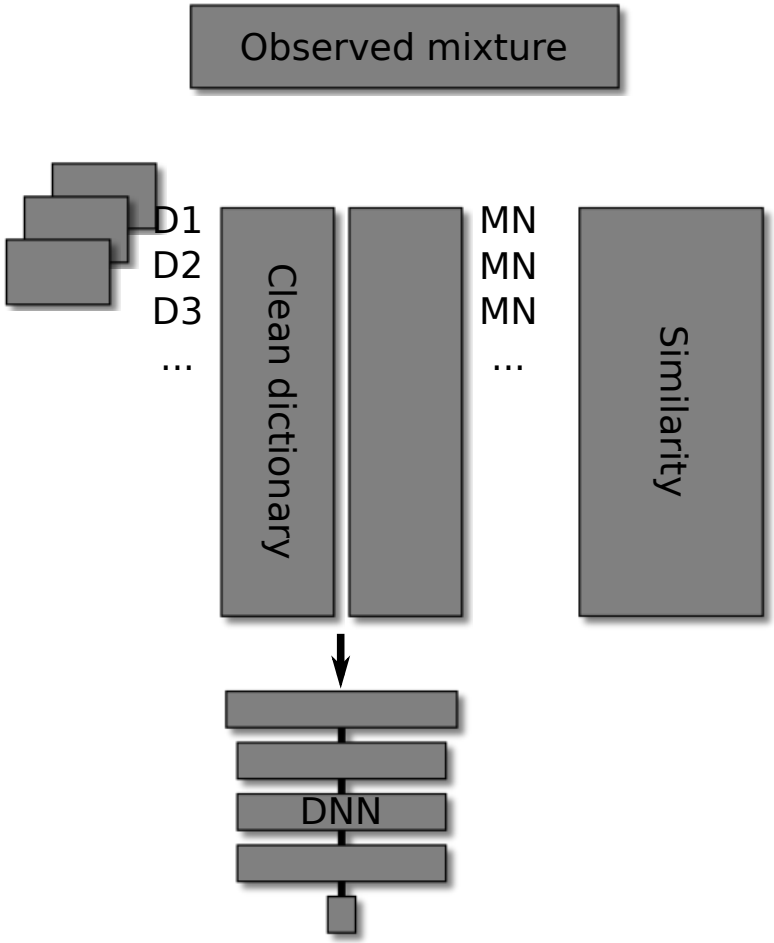
# Compare all pairs of noisy and clean chunks



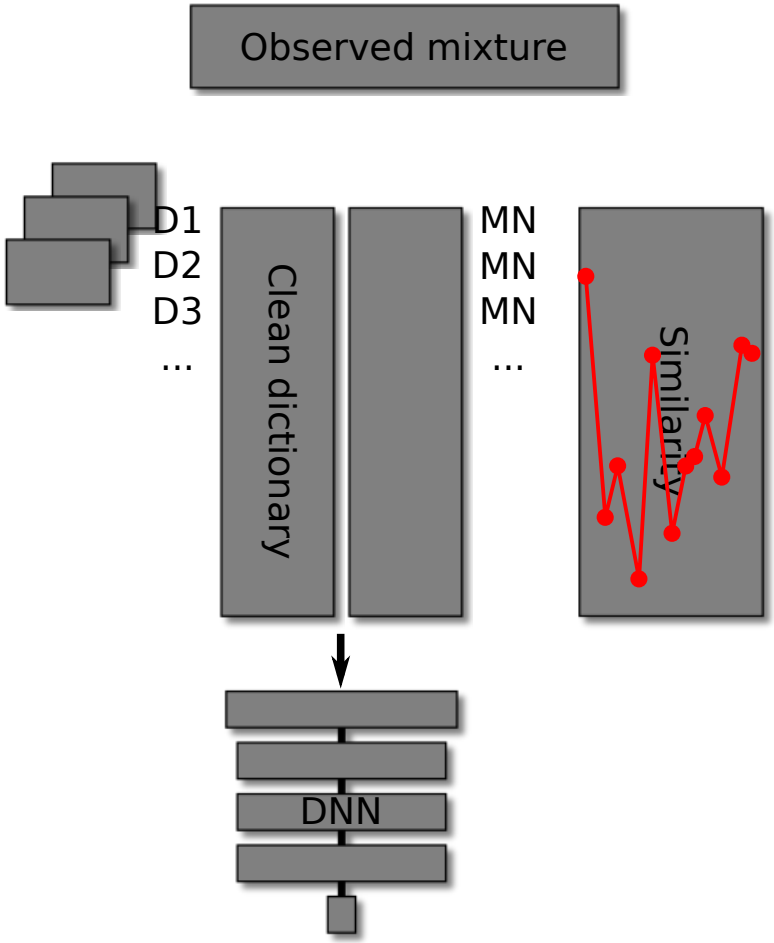
# Compare all pairs of noisy and clean chunks



# Standard Viterbi algorithm for to find optimal sequence



# Standard Viterbi algorithm for to find optimal sequence



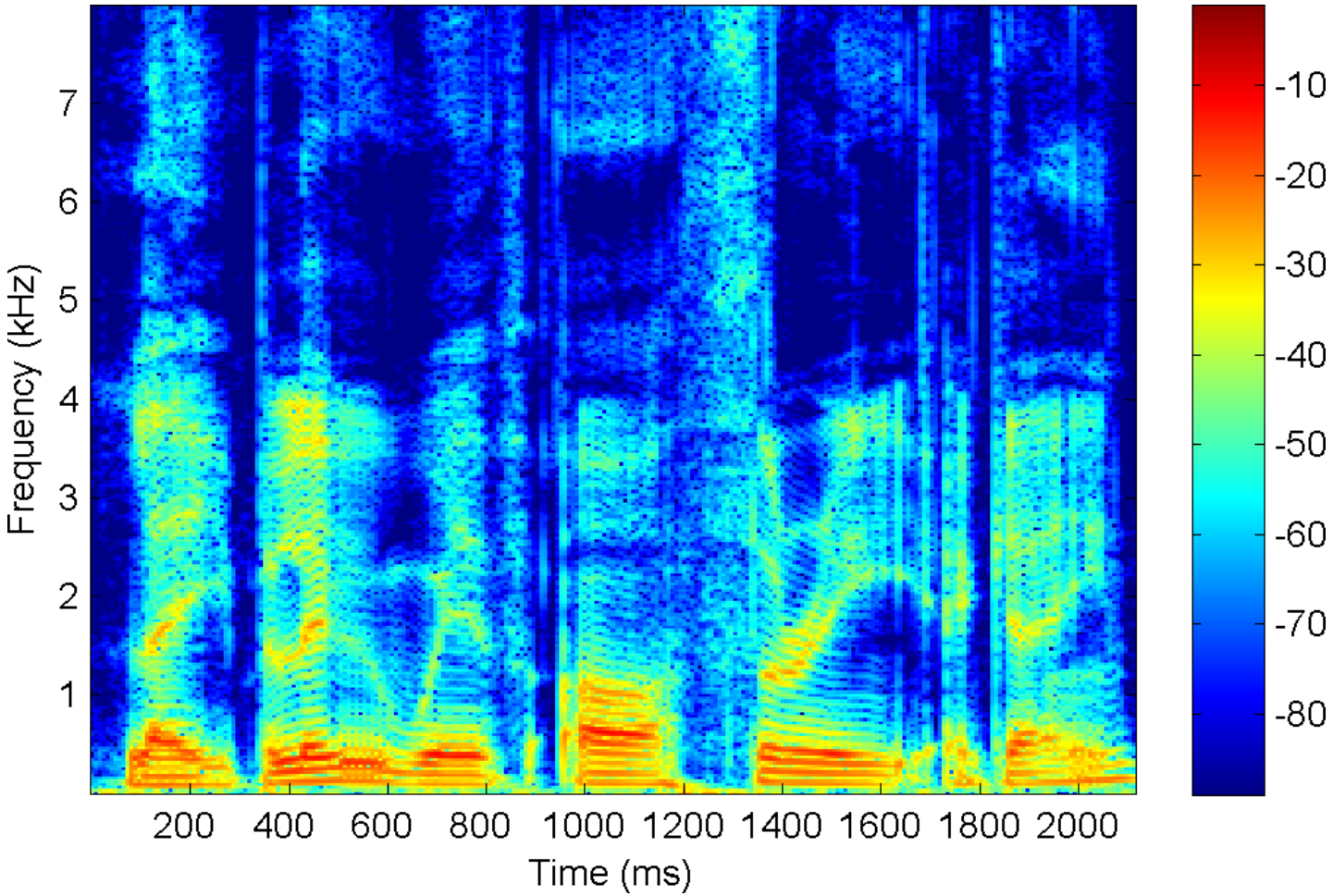
# Outline

# Experimental setup

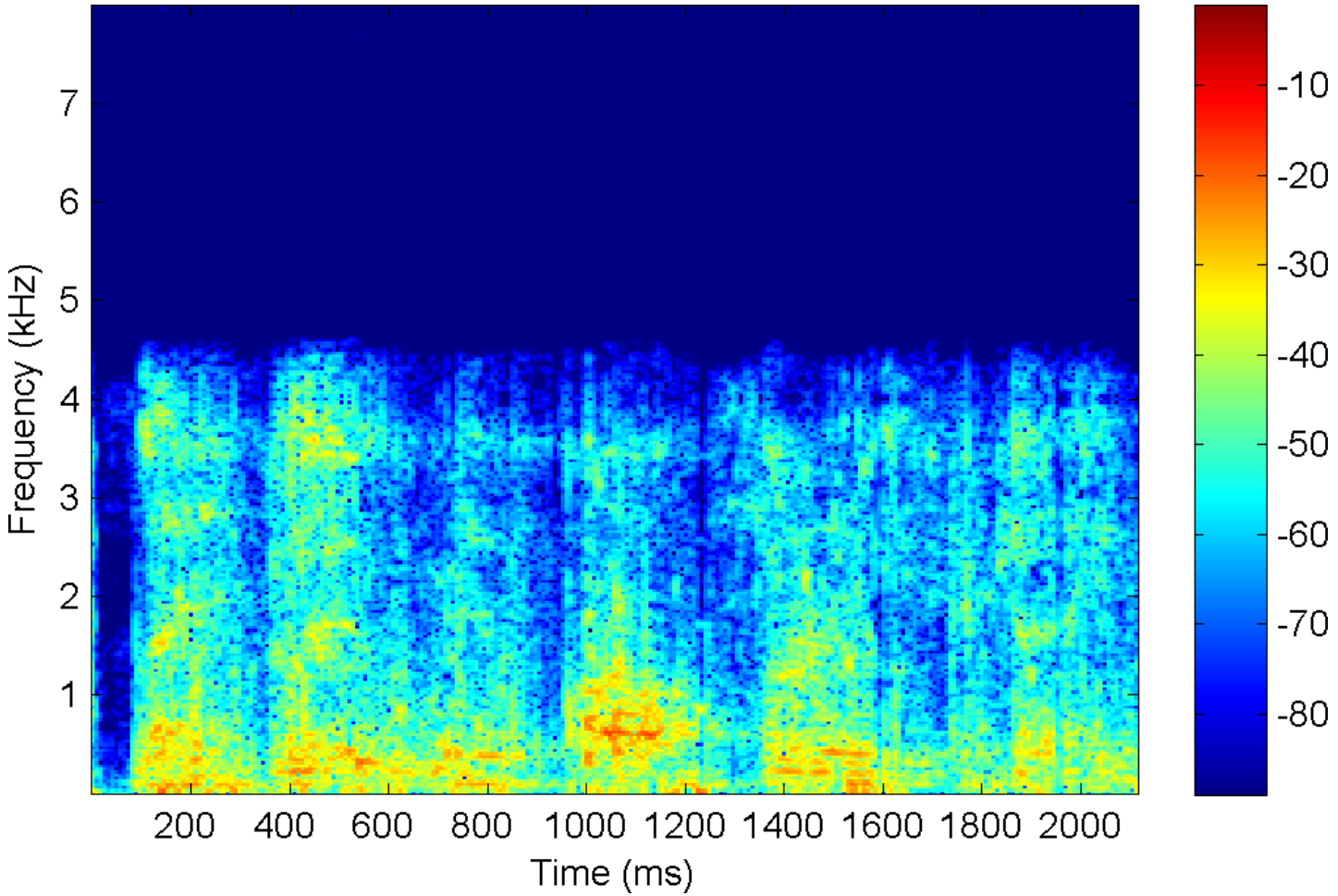
- CHiME2-GRID Corpus
  - small vocabulary task (50 words)
  - constrained grammar
- Single talker (10 minutes of training speech)
- Reverberated, encoded with Opus codec, dropped 20% of packets
- Compare against NMF bandwidth enhancement[2] (our impl.)
- Measure intelligibility and quality via listening tests (8 listeners)

[2] Dhananjay Bansal, Bhiksha Raj, and Paris Smaragdis. Bandwidth expansion of narrowband speech using non-negative matrix factorization. In *Eurospeech*, 2005

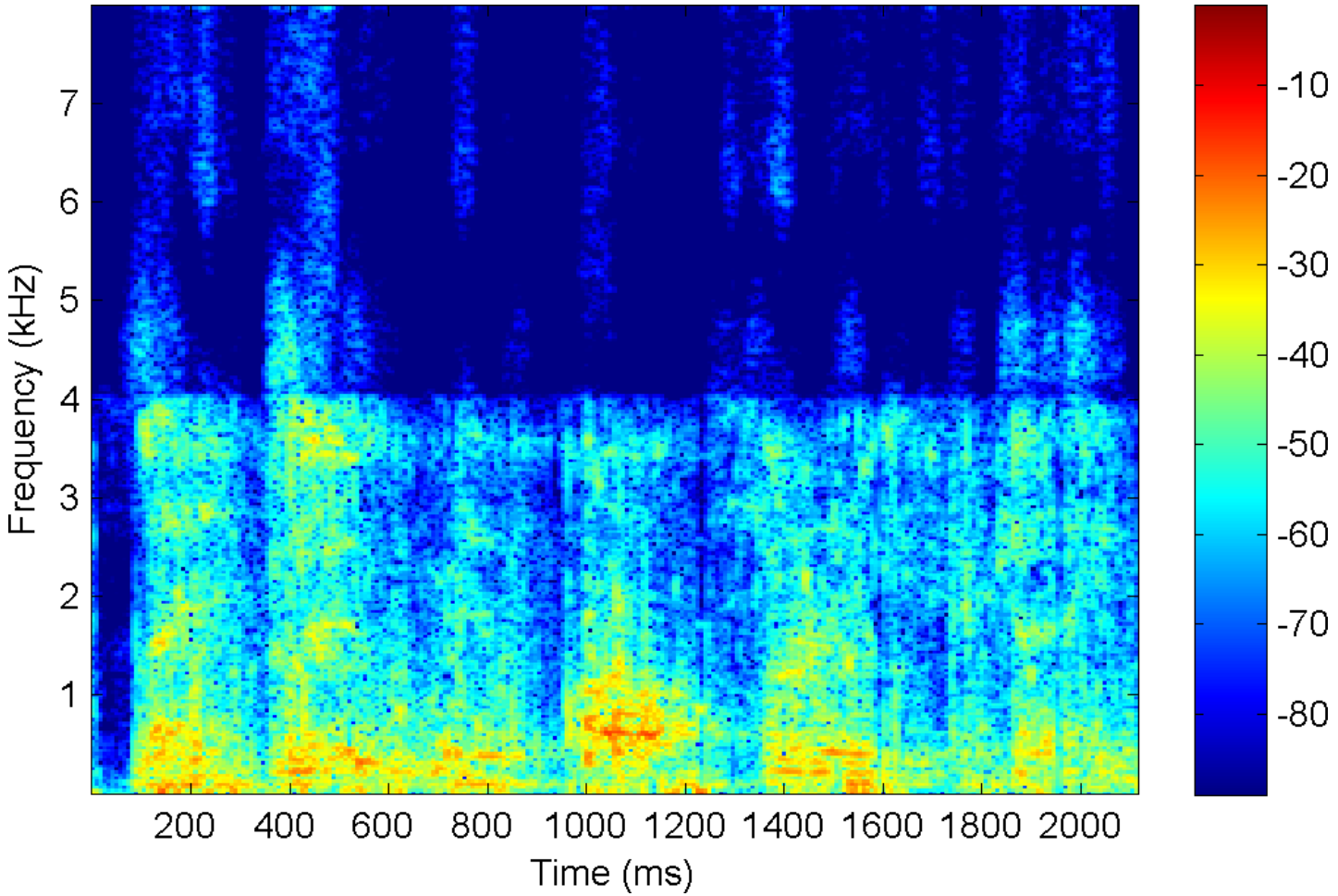
# Original clean speech



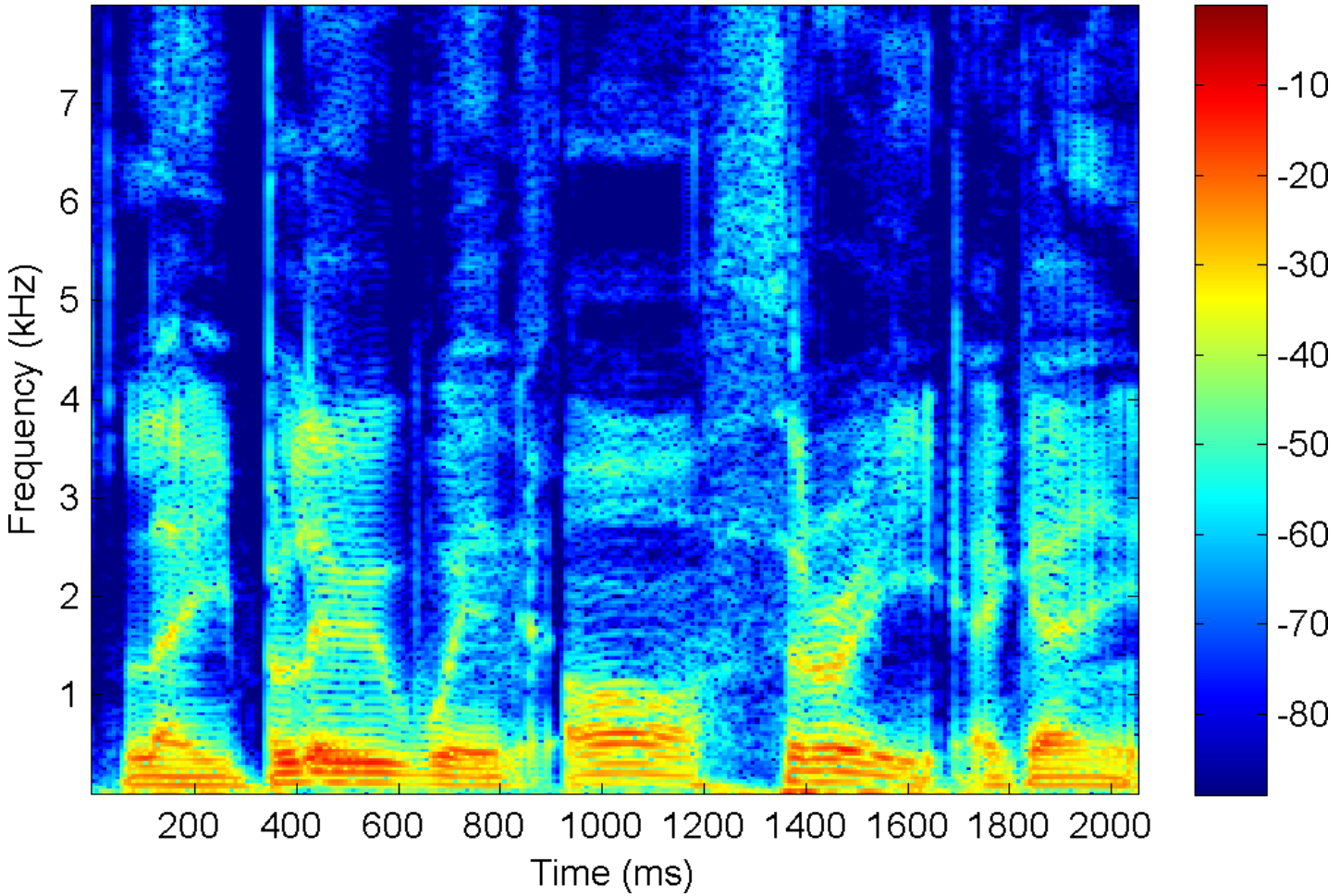
# Reverberated, compressed, 20% packet loss



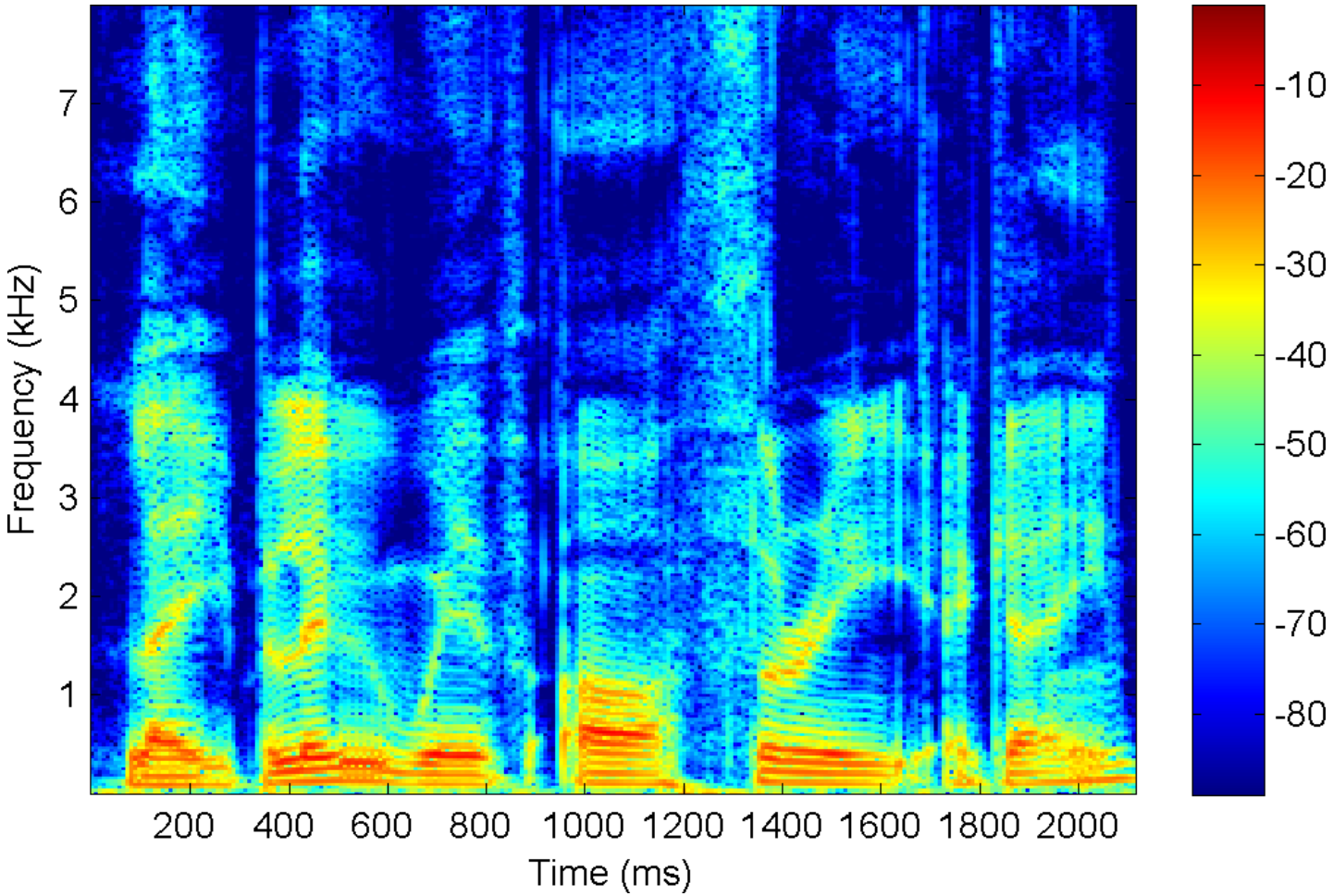
# NMF-based bandwidth expansion (our impl.)



# Concatenative resynthesis output

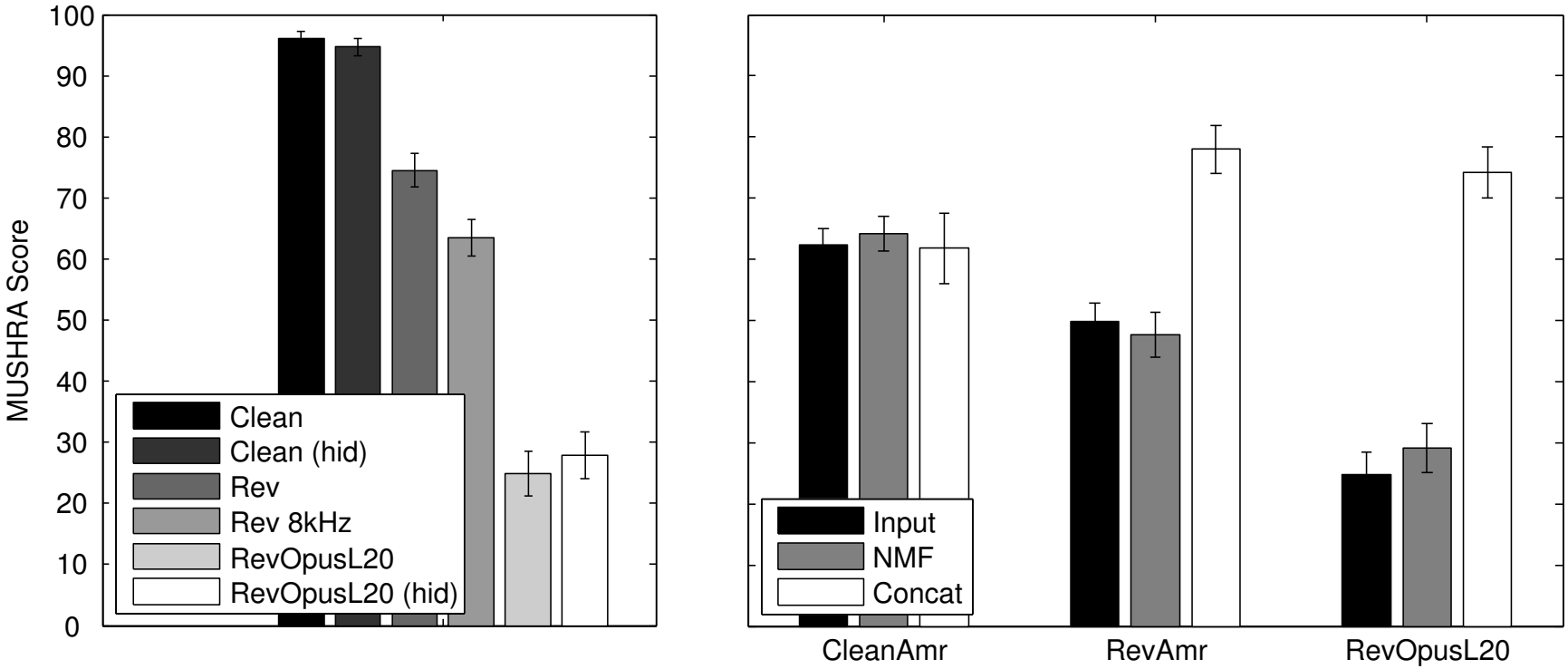


# Original clean speech

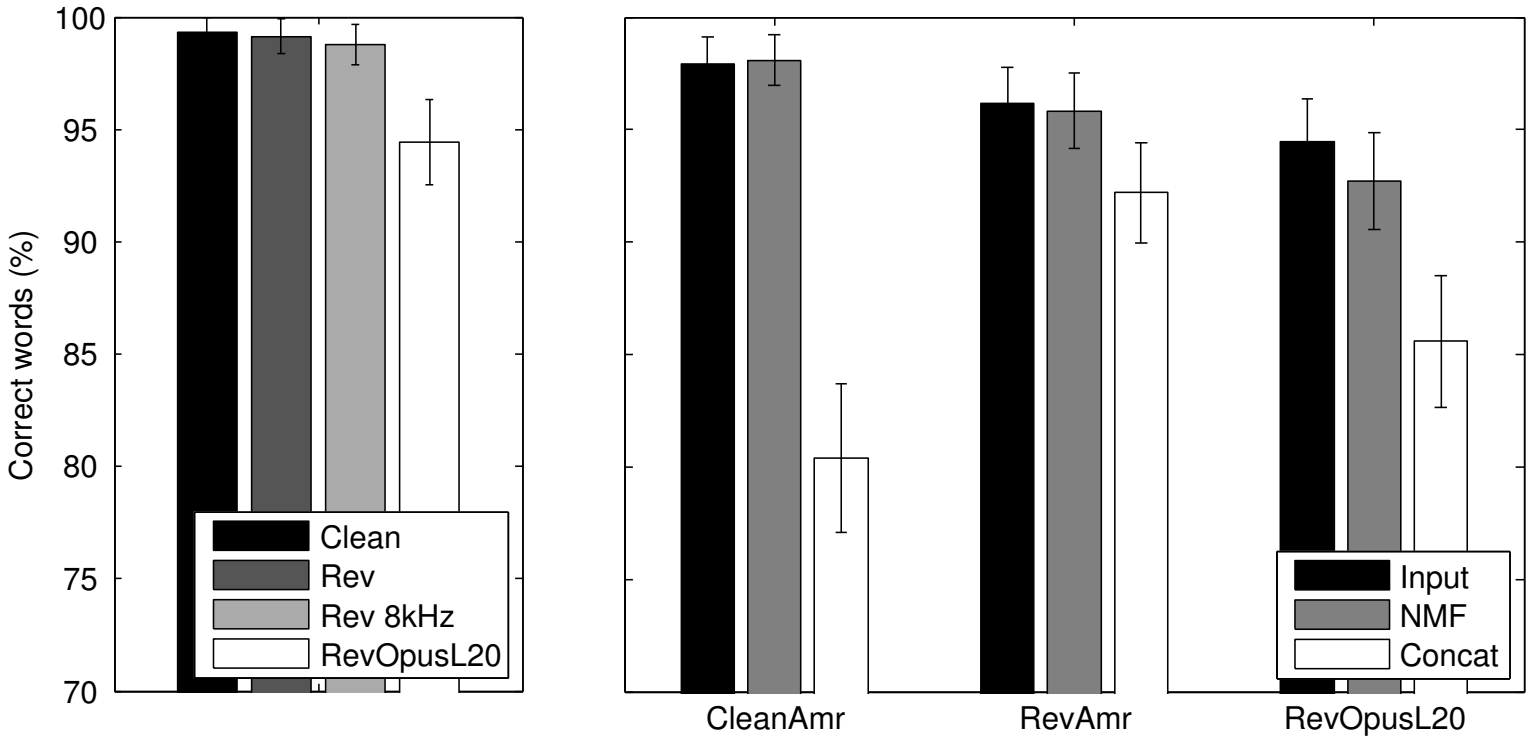


# Subjective quality is high

# 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

- 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
- Add visual features for audio-visual speech enhancement / recognition

# Thanks!

## Thanks!

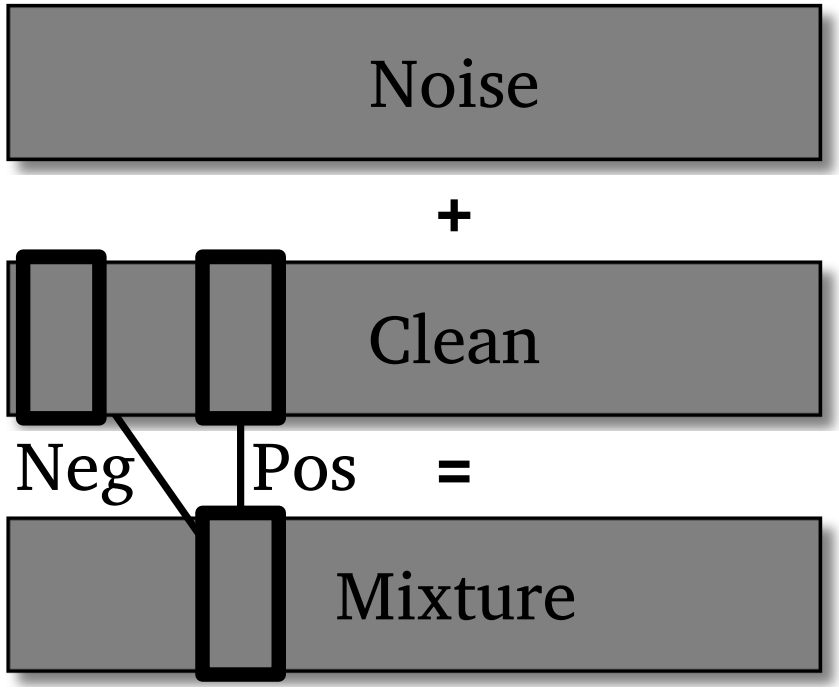
# Thanks!

**Thanks!**

Any questions?

- 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

# Train DNN on correctly and incorrectly paired chunks




Noise suppression