Combining Spatial Clustering with LSTM Speech Models for Multichannel Speech Enhancement

Zhaoheng Ni
Felix Grezes, Viet Anh Trinh, Michael Mandel

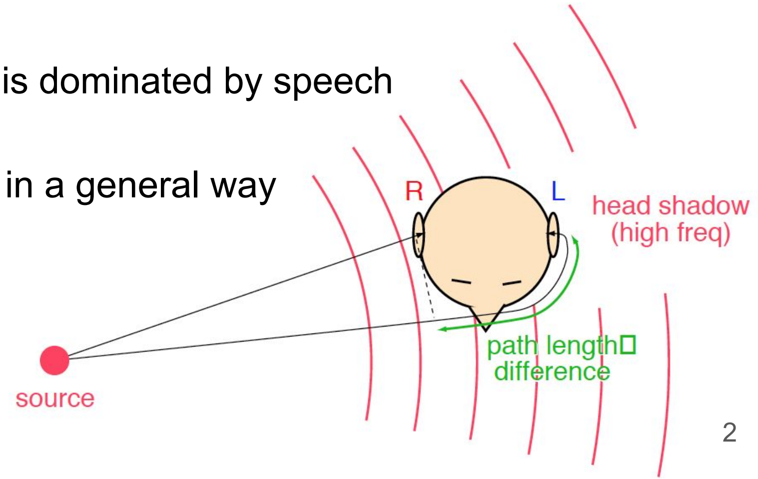
Mid-Atlantic Student Colloquium on Speech, Language and Learning, May 6th 2017.

Speech Lab @ BC-CUNY

Motivation

- **Spatial clustering** groups spectrogram points by predicted direction of arrival
  - Our system: Model-based EM Source Separation and Localization (MESSL)
  - Traditionally signal-agnostic, so doesn't take advantage of known signals

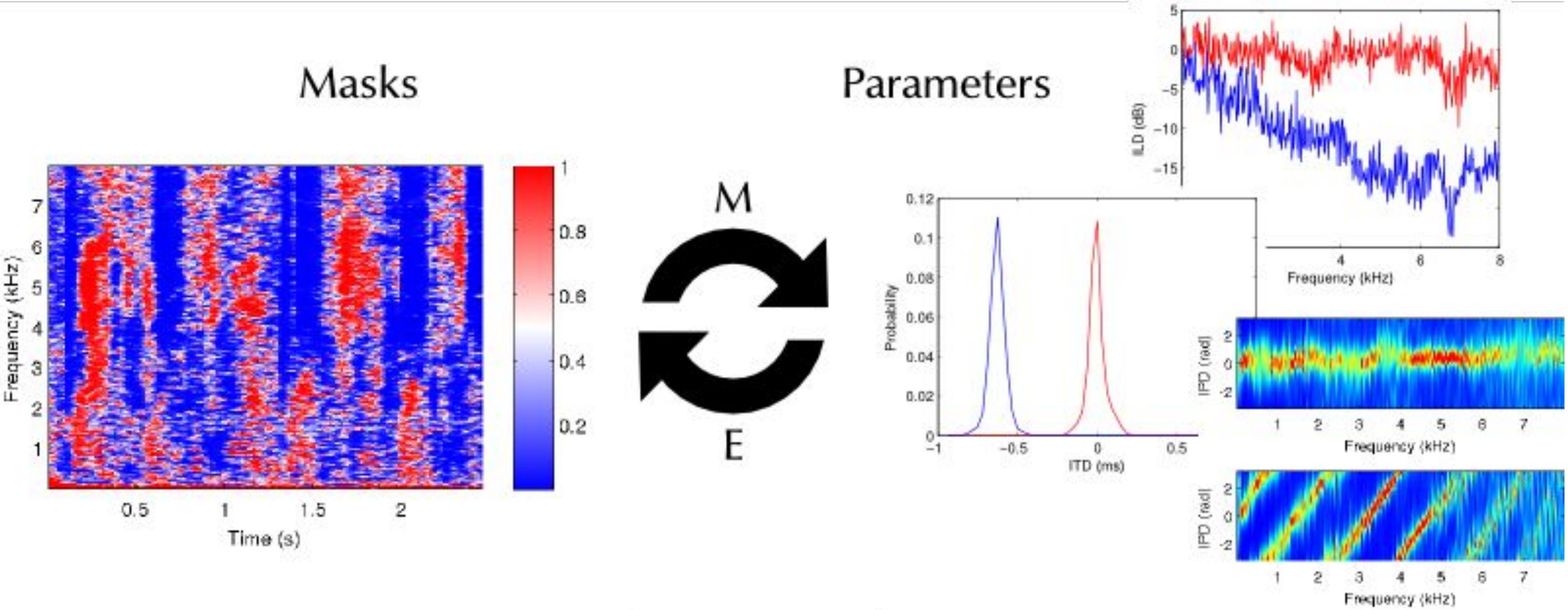
- **Deep learning**-based speech enhancement
  - Attempts to predict whether each spectrogram point is dominated by speech
  - Models signal well
  - Difficult to incorporate spatial information, especially in a general way



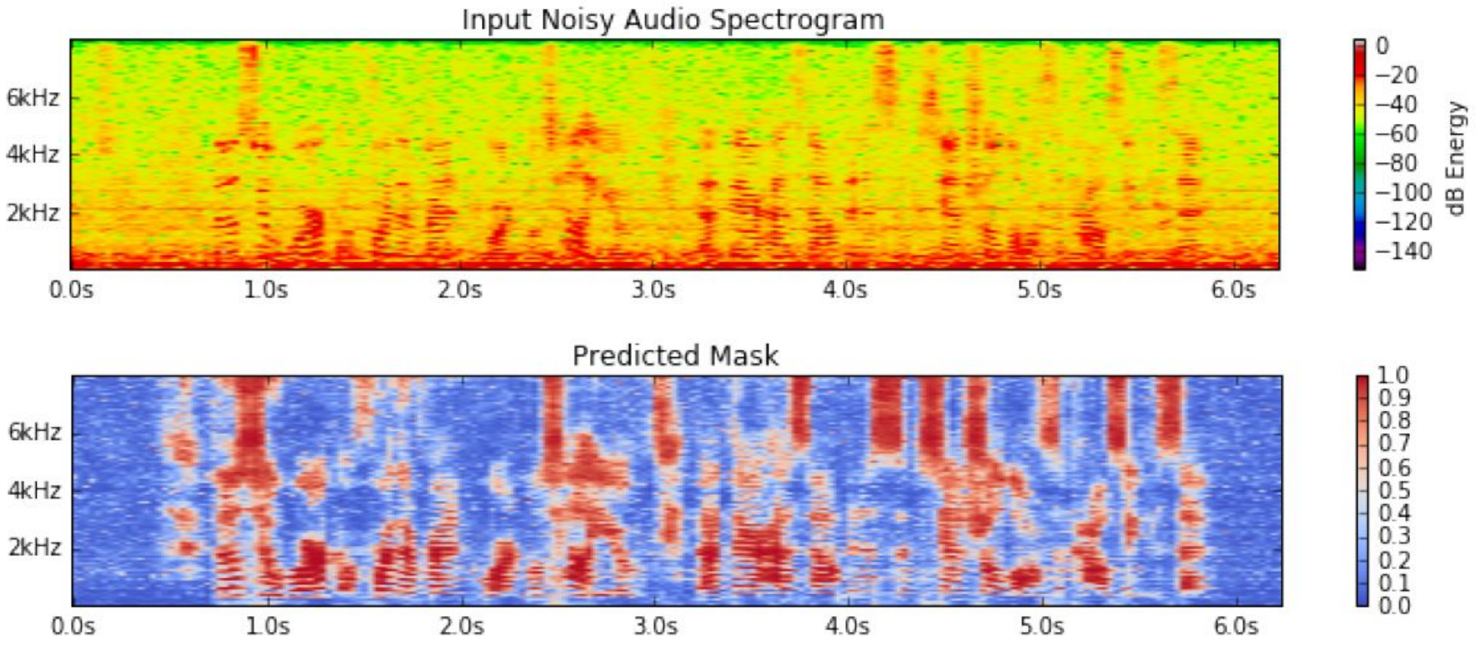
Outline

- Motivation
- Spatial Clustering (MESSL)
- LSTM Speech Enhancement Model
- Methods
- Evaluation
  - Speech quality: PESQ
  - Speech intelligibility: WER
- Results
- Conclusion
- Technical Details

Spatial Clustering (MESSL)



LSTM Speech Enhancement Model



Given a single-channel spectrogram, predict the time-frequency mask.

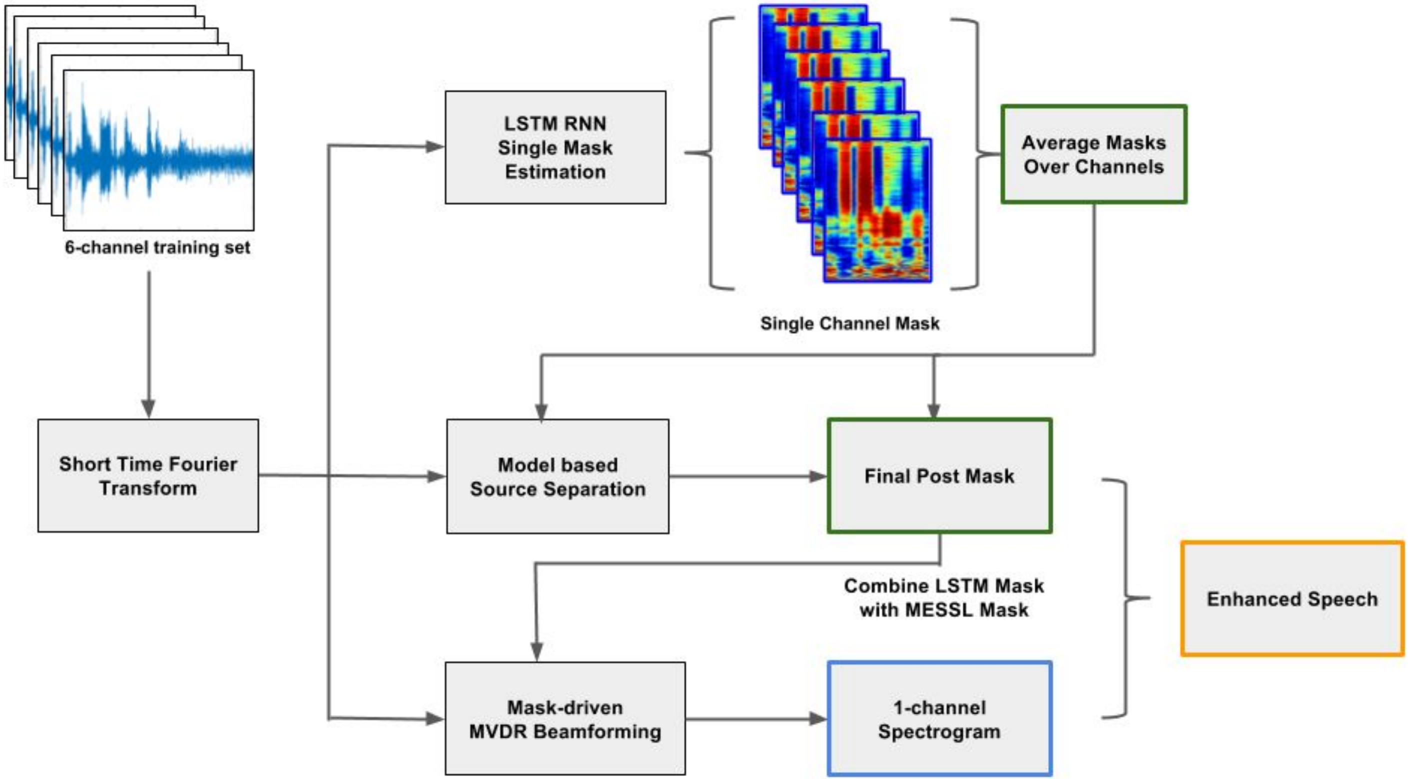
Methods

- Combine MESSL with deep learning-based single-channel signal model.
  - The model is a sequence-to-sequence Long Short-Term Memory (LSTM) neural network.

- We compare several combinations:
  1. Combining the MESSL masks with the single-channel LSTM masks.
  2. Using the LSTM masks to initialize the MESSL EM algorithm.
  3. Training an LSTM “mask cleaner” to enhance the MESSL masks.

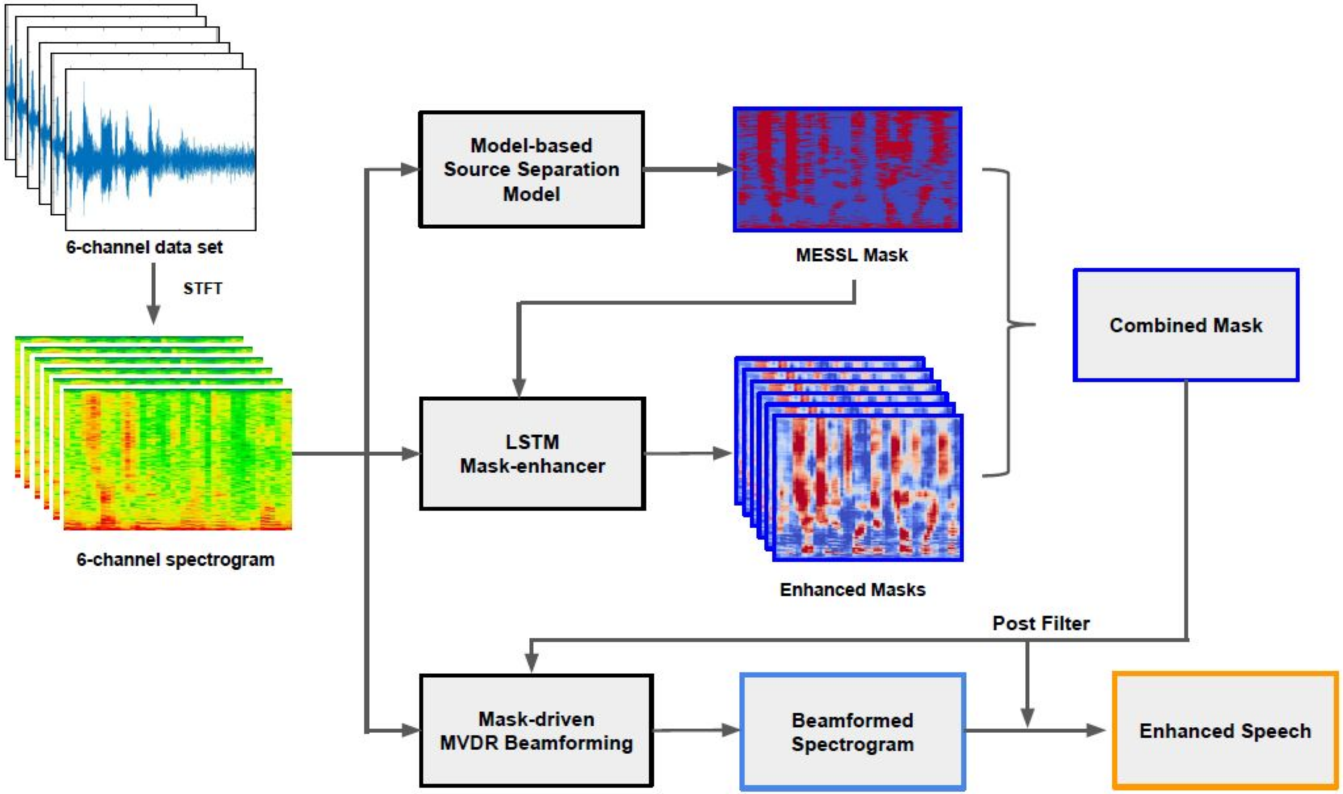
- Work done on the CHiME-3 dataset.
  - Noisy 6-channel audio, 12 speakers, 4 environments
  - ~3 hours per training, validation and testing.

Our System:



6-channel training set
LSTM RNN
Single Mask
Estimation
Single Channel Mask
Average Masks
Over Channels
Short Time Fourier
Transform
Model based
Source Separation
Final Post Mask
Combine LSTM Mask
with MESSL Mask
Enhanced Speech
Mask-driven
MVDR Beamforming
1-channel
Spectrogram

System 2: Mask Enhancer



Evaluation

- Speech quality: Perceptual Evaluation of Speech Quality (PESQ)

- Speech intelligibility: Word Error Rate (WER) as given by an ASR system trained on a different corpus.
  - Train on AMI multi-mic processed by BeamformIt (78 hours)
    - 8-mic meeting recordings: far-field, reverberant

AMI

CHiME-3

Results

Perceptual evaluation of speech quality (0-5, higher is better)

Models	validation	testing
MESSL only	1.92	1.57
LSTM only	2.51	2.42
1 - MESSL+LSTM	2.73	2.49
2 - LSTM-initialized MESSL	**2.76**	2.46
3 - LSTM Mask enhancer *	2.72	2.47

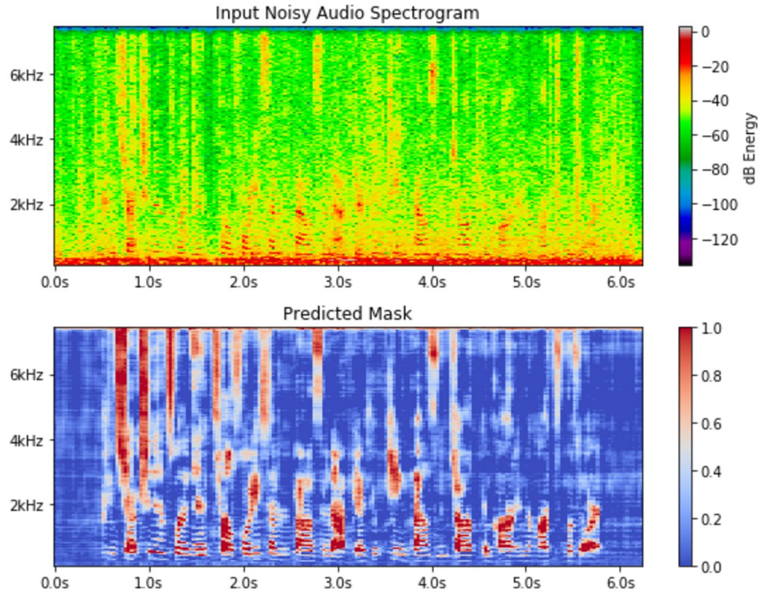
Results

Word error rate (0-100%, lower is better)

Models	validation	testing
MESSL only	26.6	43.9
LSTM only	32.9	38.9
1 - MESSL+LSTM	22.6	36.1
2 - LSTM-initialized MESSL	22.1	32.7
3 - LSTM Mask enhancer *	**19.3**	32.6

Examples

Models	isolated/dt05_bus_real/ F01_050C0103_BUS.CH5.wav
Original Noisy	
MESSL only	
LSTM only	
1 - MESSL+LSTM	
2 - LSTM-initialized MESSL	
3 - LSTM Mask enhancer	



Conclusion

Combining spatial clustering with an LSTM speech model enhances noisy audio both for speech quality and speech intelligibility.

Thanks!

Technical Details

LSTM Training Targets:

	Training Targets	Loss Functions
Ideal Amplitude Masks	$m_{ia}(\omega,t) = \lvert s(\omega,t)\rvert / \lvert y(\omega,t)\rvert$	Binary Cross Entropy
Phase Sensitive Masks	$m_{ps}(\omega,t) = \cos(\theta_{\omega,t})\frac{\lvert s(\omega,t)\rvert}{\lvert y(\omega,t)\rvert}$	Binary Cross Entropy
Magnitude Spectrum Approximation	$m_{ma}(\omega,t) = \lvert s(\omega,t)\rvert$	Mean-Squared Error
Phase-sensitive Spectrum Approximation	$m_{pa}(\omega,t) = \cos(\theta_{\omega,t})\lvert s(\omega,t)\rvert$	Mean-Squared Error