WHAMR!: Noisy and Reverberant Single-Channel Speech Separation

Matthew Maciejewski[1,2], Gordon Wichern[1], Emmett McQuinn[3], Jonathan Le Roux[1]

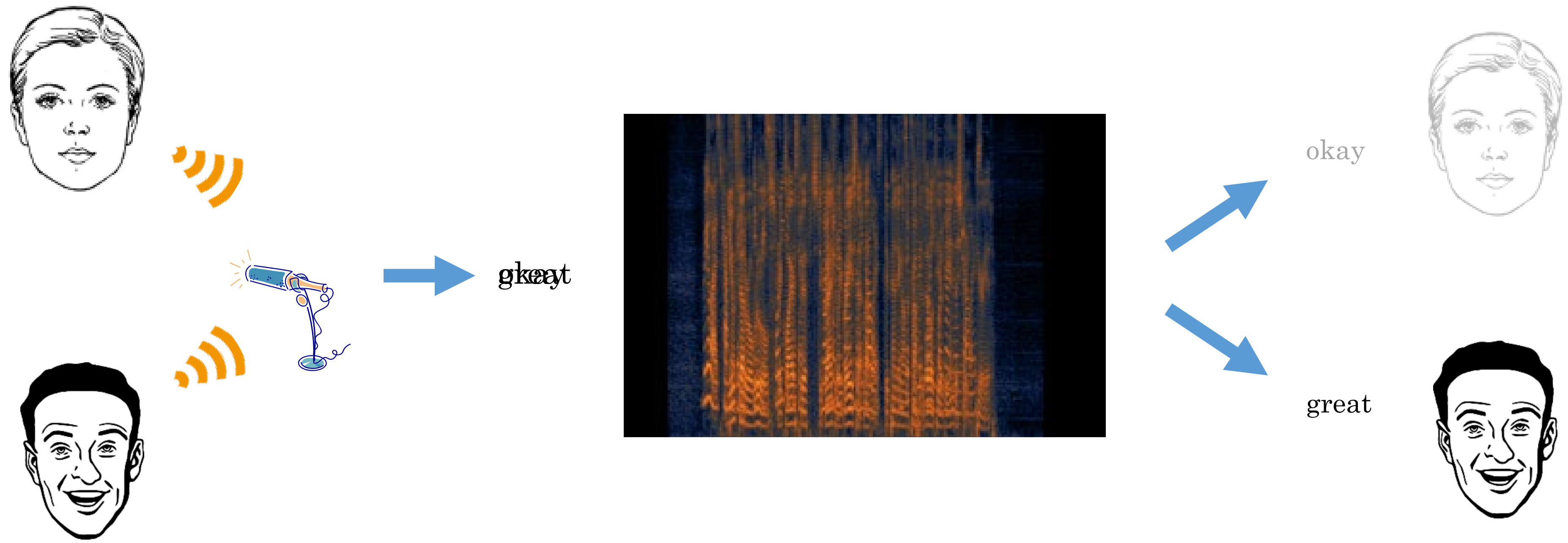
[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Johns Hopkins University, Baltimore, MD, USA
[3]whisper.ai, San Francisco, CA, USA

May 6, 2020

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

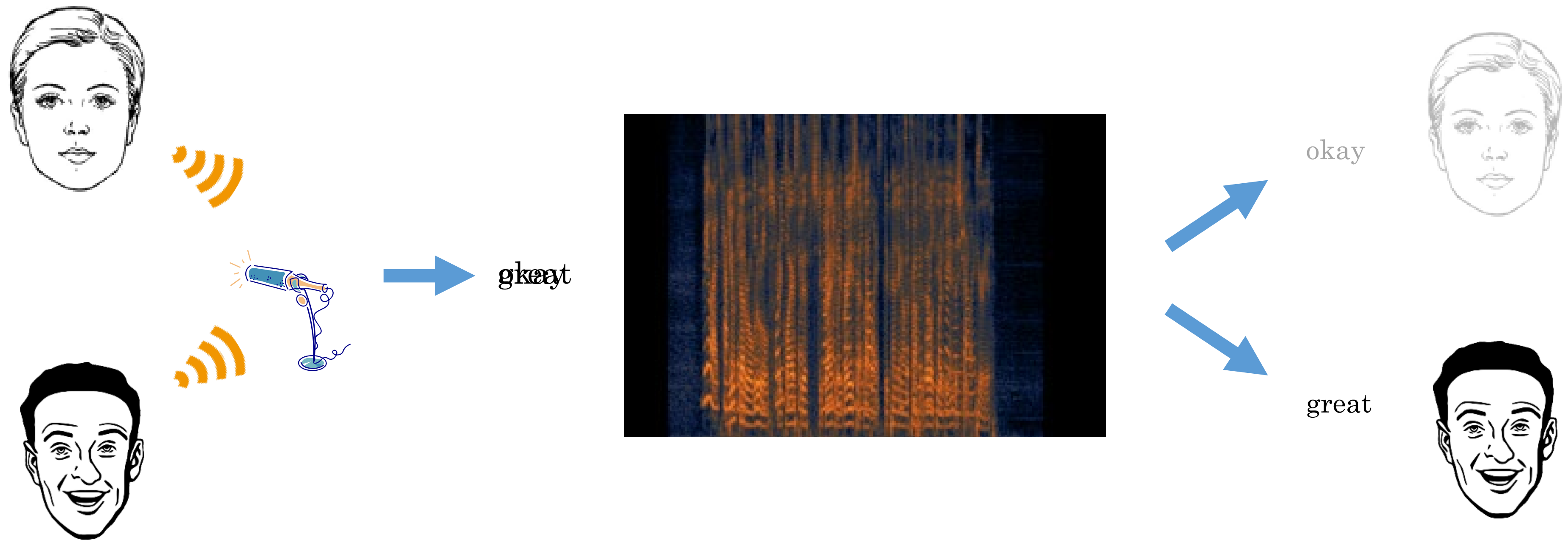
This work was performed while M. Maciejewski was an intern at MERL.

What is speech separation?

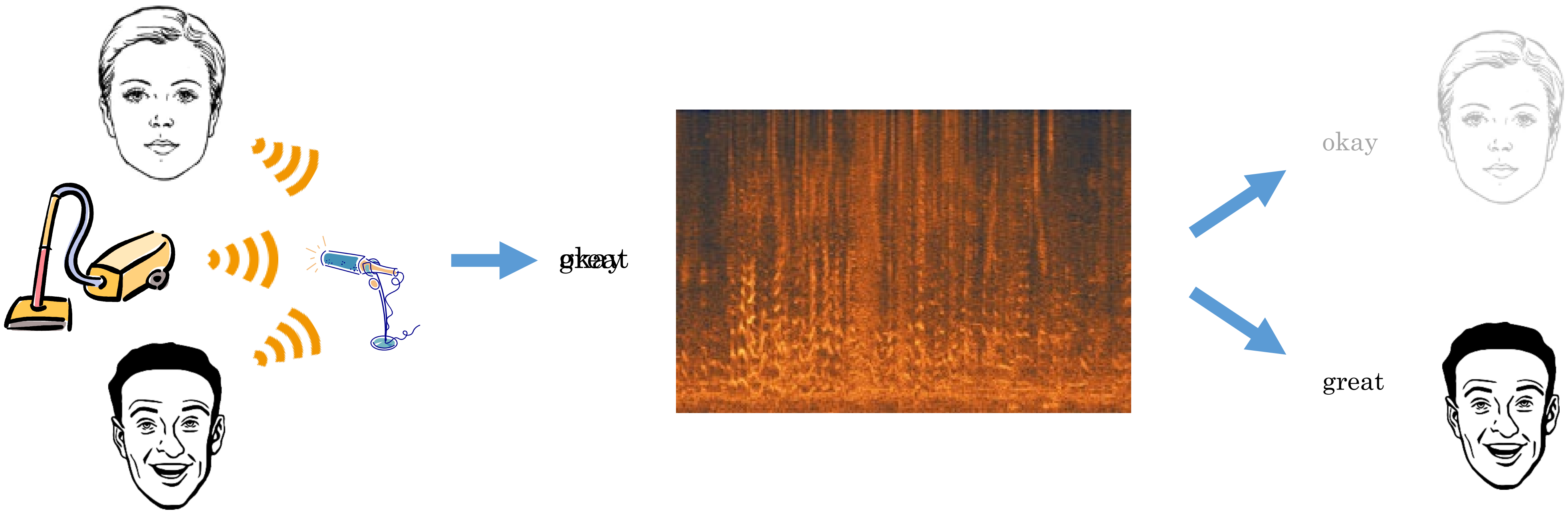


- Producing multiple single-speaker recordings from a recording of overlapped speech

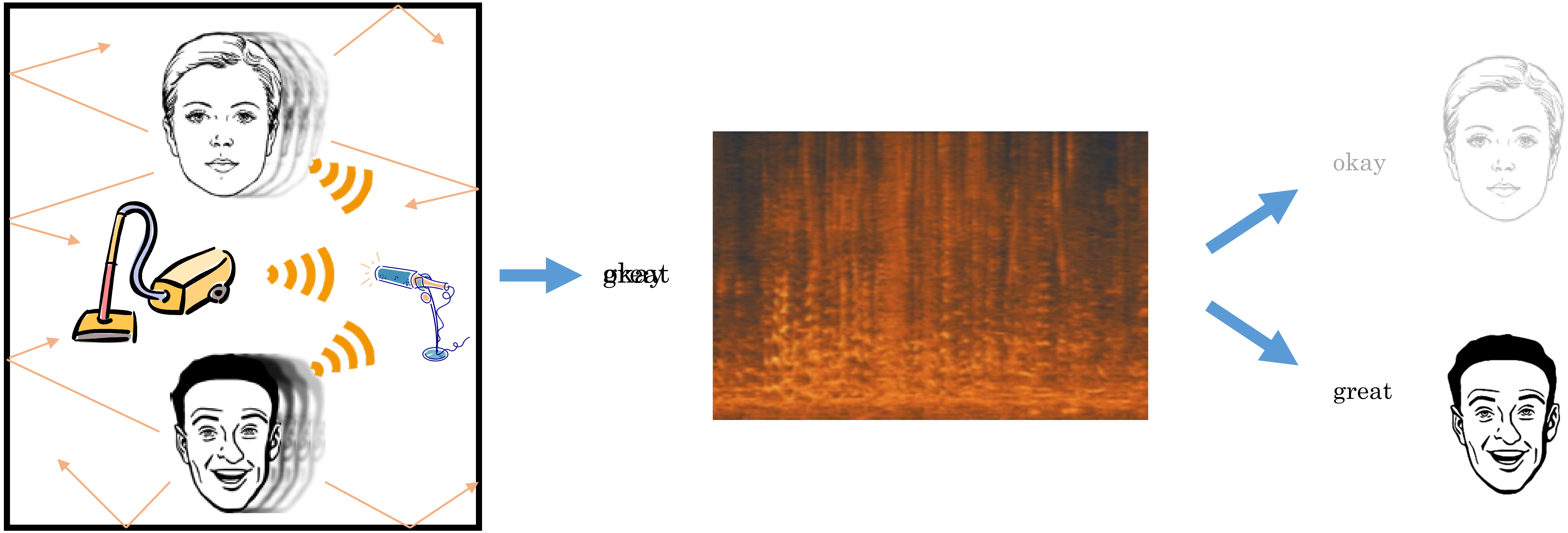
Why WHAMR!?



Why WHAMR!?



Why WHAMR!?



Pre-Existing MERL Datasets

wsj0-2mix

- Mixtures of WSJ0 corpus recordings (studio read speech)
- Standard corpus used in speech separation

WHAM!

(WSJ0 Hipster Ambient Mixtures)

- wsj0-2mix augmented with noise recorded from real environments in San Francisco
  - Noises recorded in coffee shops, restaurants, and bars

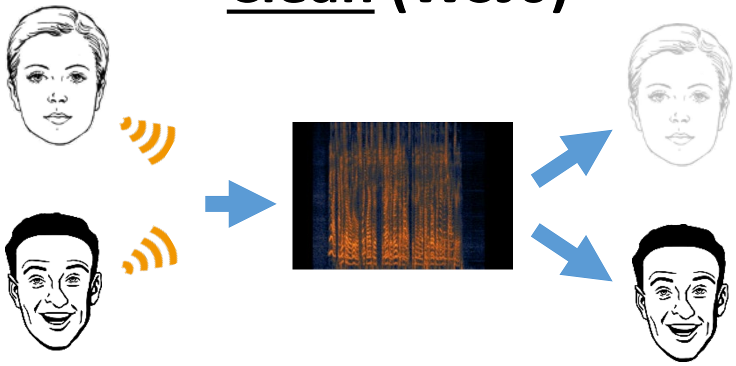
WHAMR! Dataset

- WHAM! augmented with synthetic reverberation
  - Room impulse responses generated using image-source method
  - Room parameters randomly generated to roughly match noise recordings

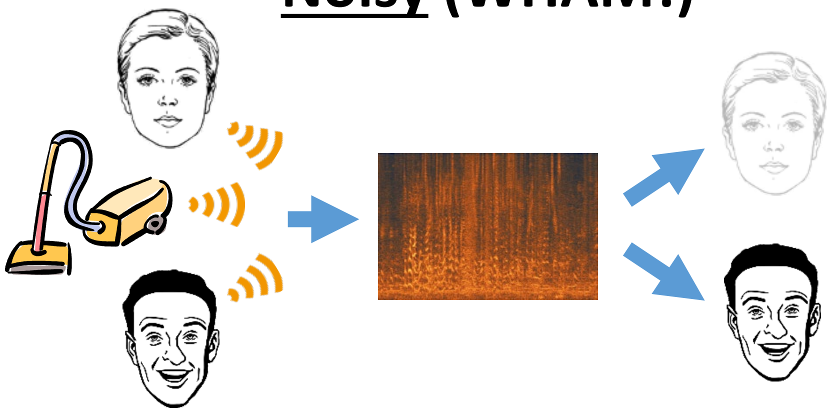
- Includes all combinations of sources, noise, and reverberation

WHAMR! Core Conditions

Clean (WSJ0)

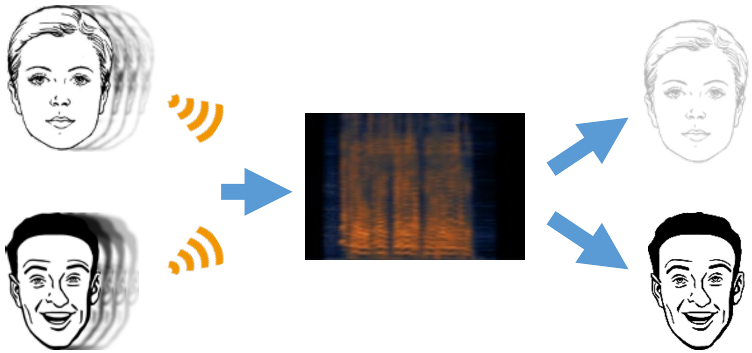


Noisy (WHAM!)

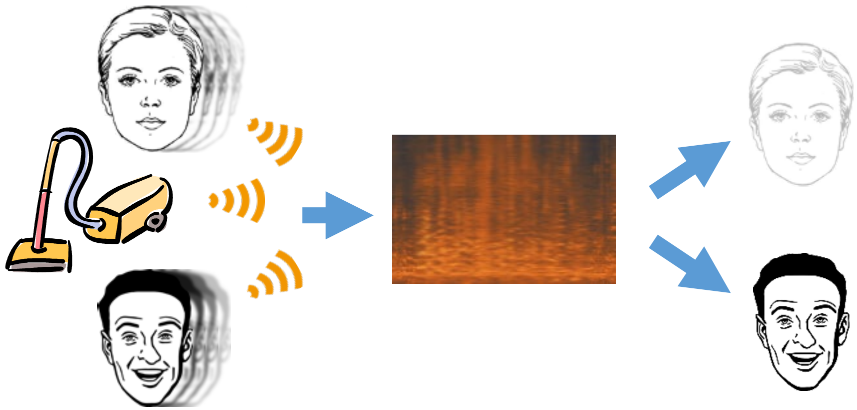


New to WHAMR!

Reverberant

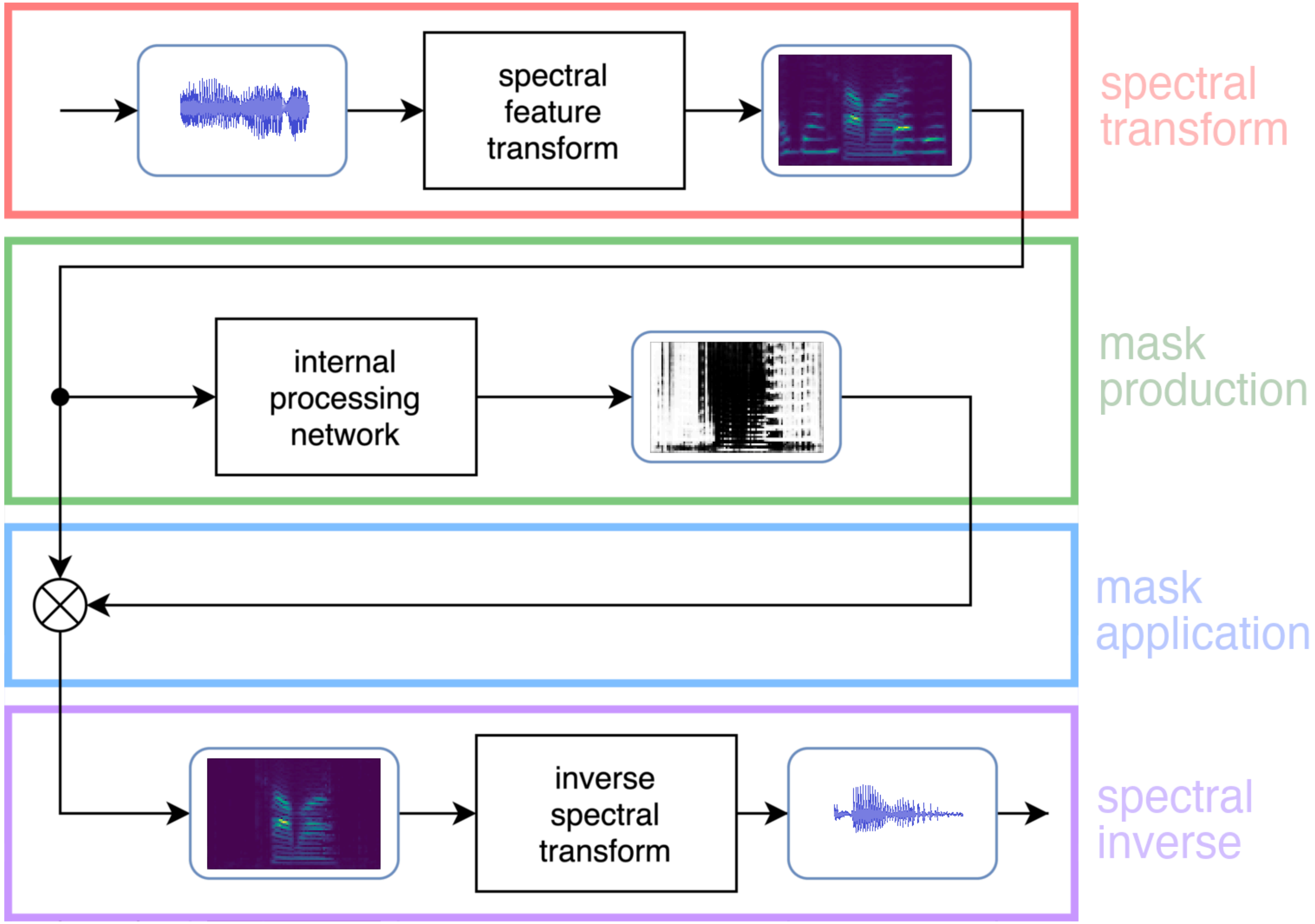


Noisy and Reverberant



Separation/Enhancement Methods

- Paired transforms between waveform and a time-frequency spectral domain
- Spectral mask is produced which suppresses interfering sources or noise/reverberation



Evaluated Model Configurations

Feature Transformations:

- Short-Time Fourier Transform (STFT)
- TasNet-style sliding-window learned basis projection

Internal Mask Production Architecture:

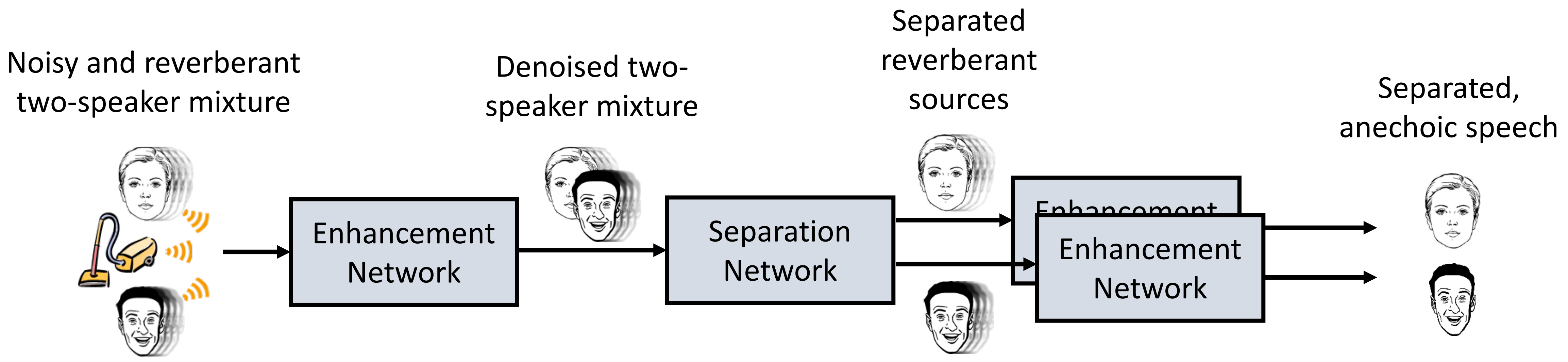
- Temporal Convolutional Network (TCN)
- Bi-directional Long Short-Term Memory (BLSTM)

All methods were trained with scale-invariant signal-to-distortions ratio (SI-SDR) loss.

SI-SDR of Core Separation Conditions using Single Model

Input			Conv-TasNet		TasNet-BLSTM	
Noise	Reverb	Input	Output	Δ	Output	Δ
		0.0	12.9	12.9	**14.2**	**14.2**
✓		−4.5	7.0	11.5	**7.5**	**12.0**
	✓	−3.3	4.3	7.6	**5.6**	**8.9**
✓	✓	−6.1	2.2	8.3	**3.0**	**9.2**

Cascaded Systems

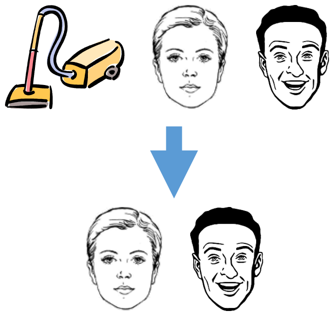


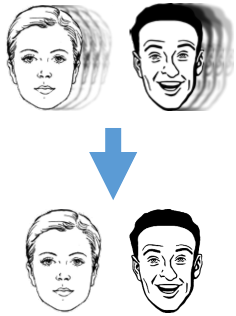
Cascaded Systems

- Pre-train separate models for each subtask
  - Separation with noisy/reverberant targets
  - Enhancement of overlapping speech
- Cascade models together

SI-SDR of Enhancement of Overlapping Speech

Net		Denoise		Dereverb	
Feature	Processor	Output	Δ	Output	Δ
Learned	TCN	10.8	9.6	7.2	3.2
Learned	BLSTM	**11.2**	**10.1**	**8.5**	**4.4**
STFT	TCN	8.4	7.2	4.0	0.0
STFT	BLSTM	9.5	8.4	5.9	1.8
Input SI-SDR:		1.2		4.0	



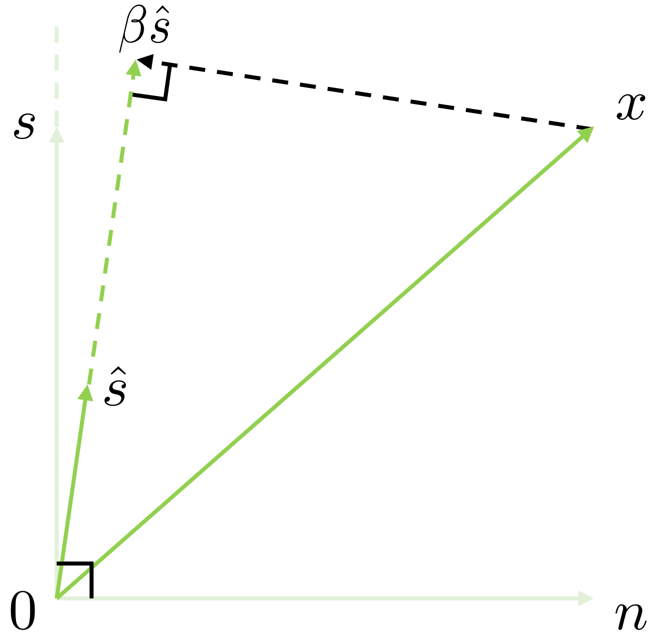


Cascaded Systems

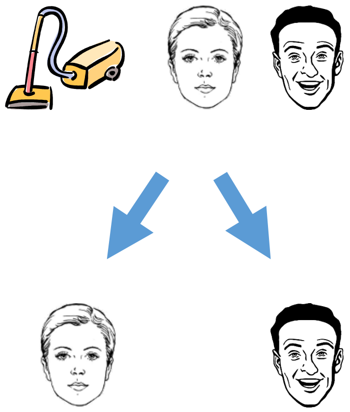
- Chain appropriately-trained models together, with rescale factor:

$$\beta(\hat{s}|x) = \frac{\langle x, \hat{s} \rangle}{\|\hat{s}\|^2}$$

  - Scale so residual is orthogonal to estimated source
  - Necessary due to scale-invariant loss.

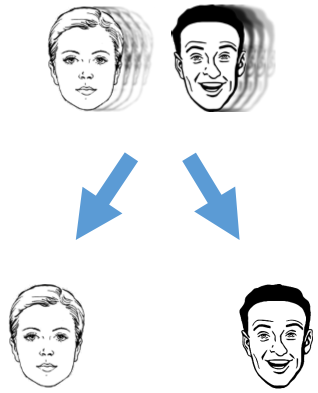


SI-SDR of Noisy Separation with Cascaded Models



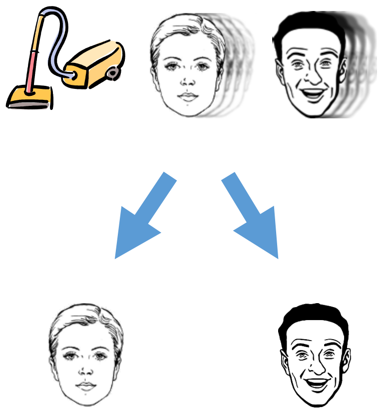
System		SI-SDR	
Pre-Enh. Removes	Separate Speech while Removing	Output	Δ
$\times$	noise	7.5	12.0
noise	–	**8.1**	**12.6**
Input SI-SDR:		-4.5	

SI-SDR of Reverberant Separation with Cascaded Models



System			SI-SDR	
Pre-Enh. Removes	Separate Speech while Removing	Post-Enh. Removes	Output	Δ
$\times$	rev.	$\times$	5.6	8.9
rev.	–	$\times$	6.4	9.7
$\times$	–	rev.	**6.6**	**9.9**
Input SI-SDR:			-3.3	

SI-SDR of Noisy and Reverberant Separation with Cascaded Models

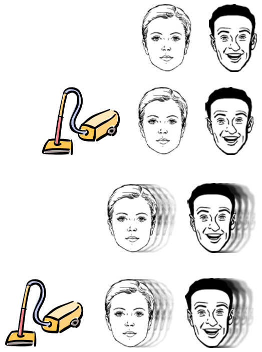


System			SI-SDR	
Pre-Enh. Removes	Separate speech while removing	Post-Enh. Removes	Output	Δ
$\times$	noise, rev.	$\times$	3.0	9.2
noise	rev.	$\times$	3.5	9.7
noise, rev.	–	$\times$	3.6	9.7
rev.	noise	$\times$	3.7	9.8
$\times$	noise	rev.	3.7	9.8
noise	–	rev.	**4.0**	**10.1**
Input SI-SDR:			-6.1	

Tuned Cascaded Systems

- Additional training epochs of full end-to-end system

SI-SDR of Tuned Cascaded Systems



Input			Best System w/o Tuning		Tuned	
Noise	Reverb	Input	Output	Δ	Output	Δ
		0.0	14.2	14.2	–	–
✓		−4.5	8.1	12.6	8.3	12.9
	✓	−3.3	6.6	9.9	7.0	10.3
✓	✓	−6.1	4.0	10.1	4.7	10.8

Conclusions

- We introduced a new speech separation dataset featuring added noise and reverberation.
- Systems with learned basis features and BLSTM processing outperform systems with STFT features and TCN processing.
- Splitting separation into subtasks of pre-separation denoising, reverberant separation, and post-separation dereverberation improves performance.

Data and creation scripts available at: http://wham.whisper.ai/

MITSUBISHI
ELECTRIC

Changes for the Better