

A Deep CASA Approach to Talker-independent Speaker Separation

DeLiang Wang
(joint with Yuzhou Liu)

Perception & Neurodynamics Lab
Ohio State University

Outline

- **Introduction**
 - The cocktail party problem
 - Computational auditory scene analysis (CASA)
 - Speech separation as DNN based mask estimation
- **Talker-dependent speaker separation**
 - Human listener test
- **Talker-independent speaker separation**
 - Deep CASA approach

Real-world audition



What?

- Speech
 - message
 - speaker
 - age, gender, linguistic origin, mood, ...
- Music
- Car passing by

Where?

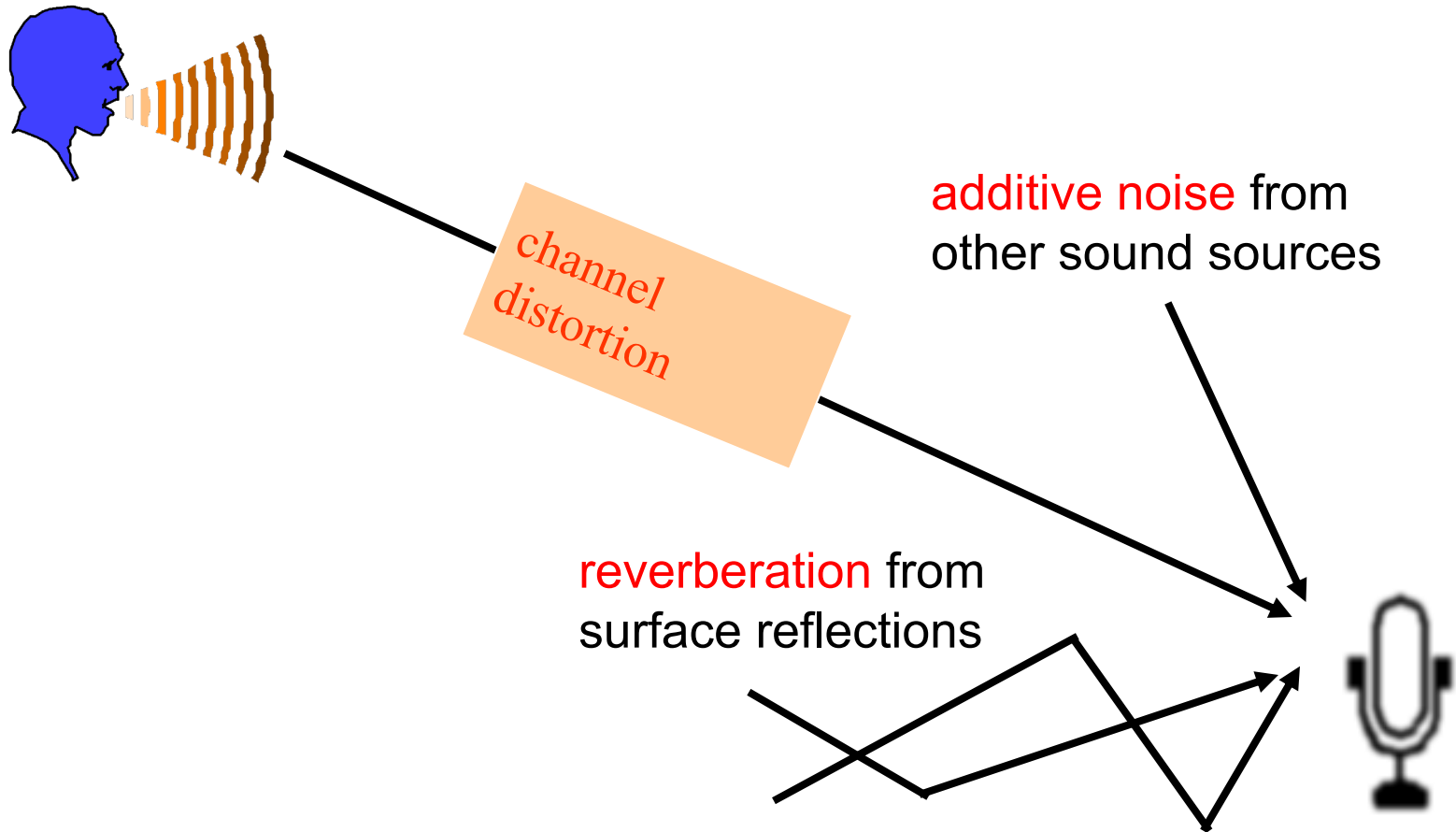
- Left, right, up, down
- How close?

Channel characteristics

Environment characteristics

- Room reverberation
- Ambient noise

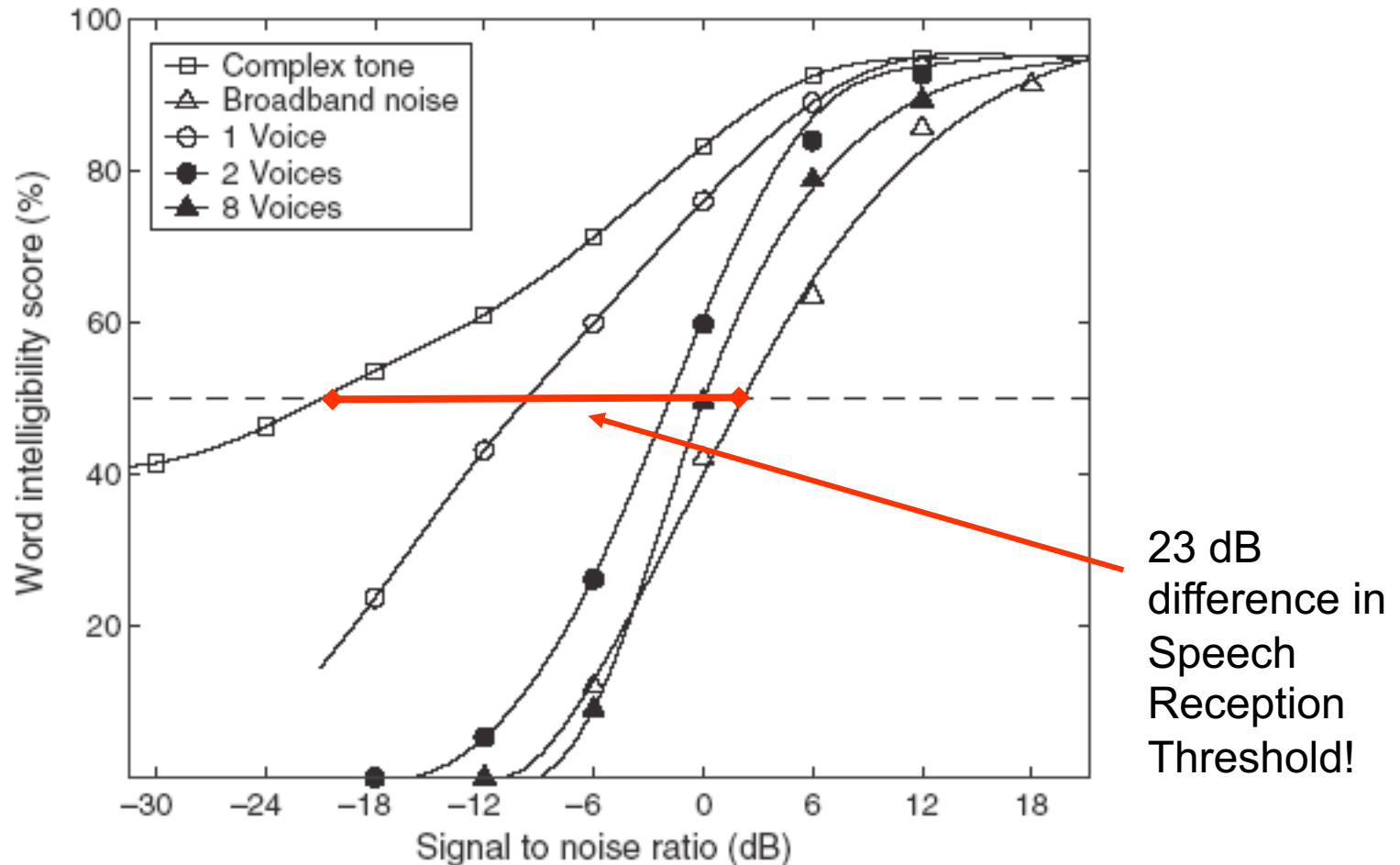
Sources of intrusion and distortion



Cocktail party problem

- **Term coined by Cherry**
 - “One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it ‘the cocktail party problem.’ No machine has been constructed to do just that.” (Cherry, 1957)
- **Speech separation problem**
 - Speech enhancement: speech-nonspeech separation
 - Speaker separation: multi-talker separation

Human performance in different interferences



Source: Wang & Brown (2006)

Auditory scene analysis (ASA)

- **Listeners are capable of parsing an acoustic scene (a sound mixture) to form a mental representation of each sound source – stream – in the perceptual process of auditory scene analysis (Bregman, 1990)**
 - From acoustic events to perceptual streams
- **Two conceptual processes of ASA:**
 - **Segmentation.** Decompose the acoustic mixture into sensory elements (segments)
 - **Grouping.** Combine segments into streams, so that segments in the same stream originate from the same source

Simultaneous organization

- **Simultaneous organization groups sound components that overlap in time. ASA cues for simultaneous organization:**
 - Proximity in frequency (spectral proximity)
 - Common periodicity
 - Common spatial location
 - Common onset (and to a lesser degree, common offset)
 - Common temporal modulation
 - Amplitude modulation
 - Frequency modulation

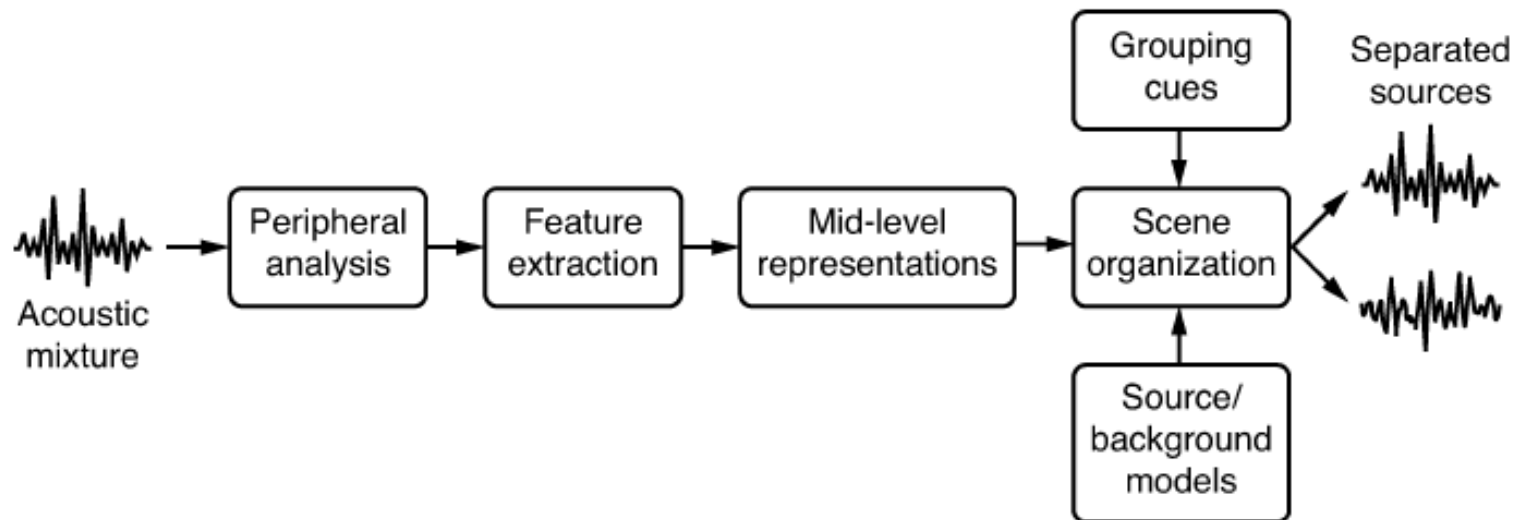


Sequential organization

- **Sequential organization groups sound components across time. ASA cues for sequential organization:**
 - Proximity/continuity in time and frequency
 - Common spatial location; more generally, spatial continuity
 - Smooth pitch contour
 - Rhythmic structure

Computational auditory scene analysis

- **Computational auditory scene analysis (CASA)**
approaches sound separation based on ASA principles
 - Feature based approaches
 - Model based approaches



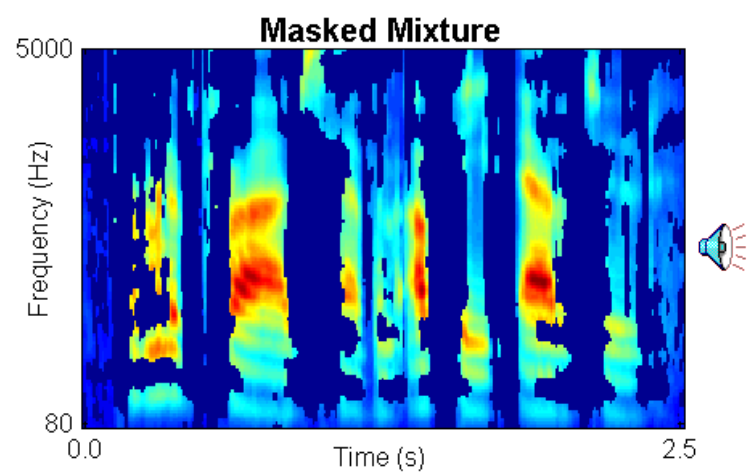
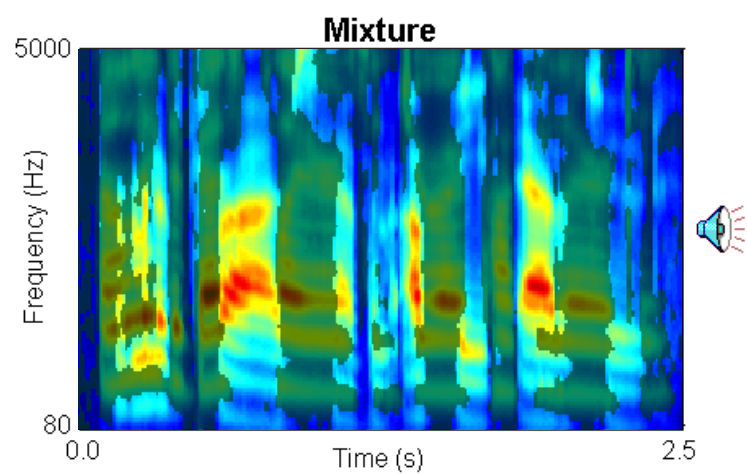
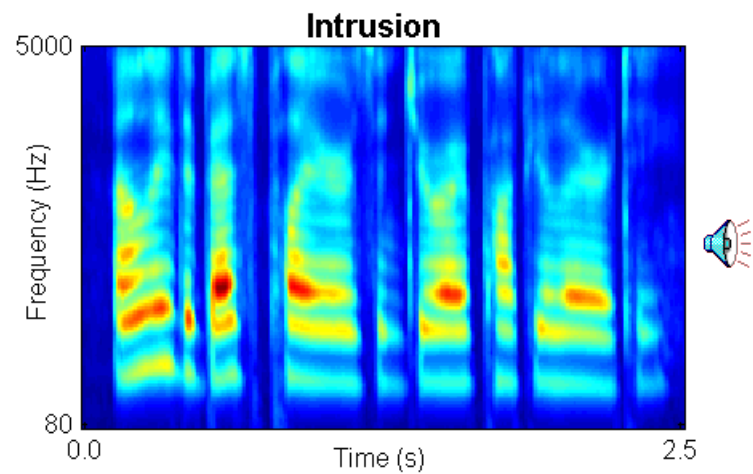
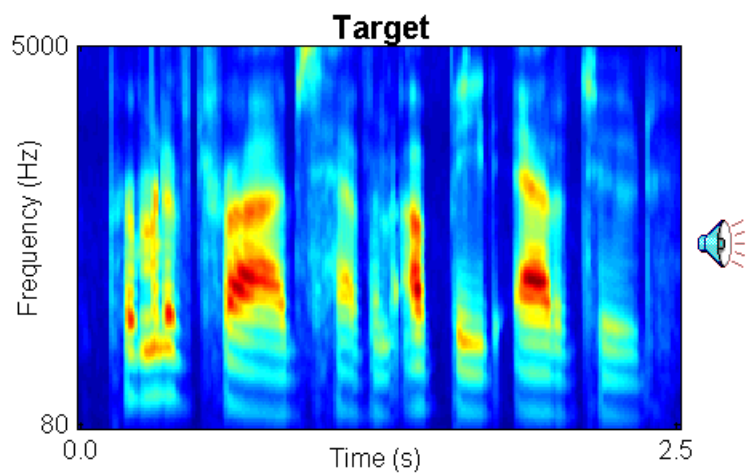
Ideal binary mask as a separation goal

- **Motivated by the auditory masking phenomenon and auditory scene analysis, we suggested the ideal binary mask as a main goal of CASA (Hu & Wang, 2001; 2004)**
- **The idea is to retain parts of a mixture where the target sound is stronger than the acoustic background, and discard the rest**
- **The definition of the ideal binary mask (IBM)**

$$IBM(t, f) = \begin{cases} 1 & \text{if } SNR(t, f) \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

- θ : A local SNR criterion (LC) in dB, which is typically chosen to be 0 dB
- Optimal SNR: Under certain conditions the IBM with $\theta = 0$ dB is the optimal binary mask in terms of SNR gain (Li & Wang, 2009)
- Maximal articulation index (AI) in a simplified version (Loizou & Kim, 2011)
- It does not actually separate the mixture

IBM illustration

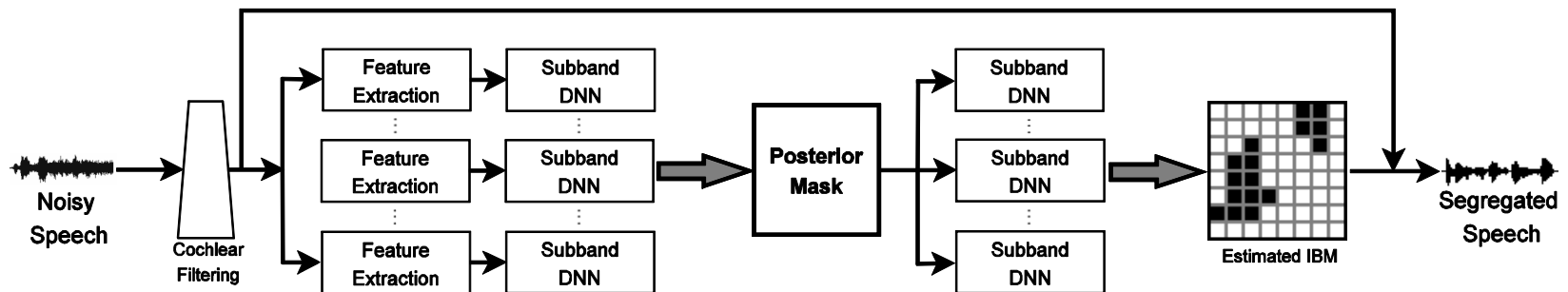


Subject tests of ideal binary masking

- **IBM separation leads to dramatic speech intelligibility improvements**
 - Improvement for stationary noise is above 7 dB for normal-hearing (NH) listeners (Brungart et al.'06; Li & Loizou'08; Ahmadi et al.'13; Chen'16), and above 9 dB for hearing-impaired (HI) listeners (Anzalone et al.'06; Wang et al.'09)
 - Improvement for modulated noise is significantly larger than for stationary noise
- **With the IBM as the goal, the speech separation problem becomes a binary classification problem**
 - This new formulation opens the problem to a variety of pattern classification methods

Speech separation as DNN based mask estimation

- **Y. Wang and Wang (2013) first introduced deep neural networks (DNNs) to address the speech separation problem**
 - DNN is used for as a subband classifier and classification aims to estimate the IBM



- **DNN based classification produced substantial speech intelligibility improvements for HI (and NH) listeners, for the first time (Healy et al.'13)**

Outline

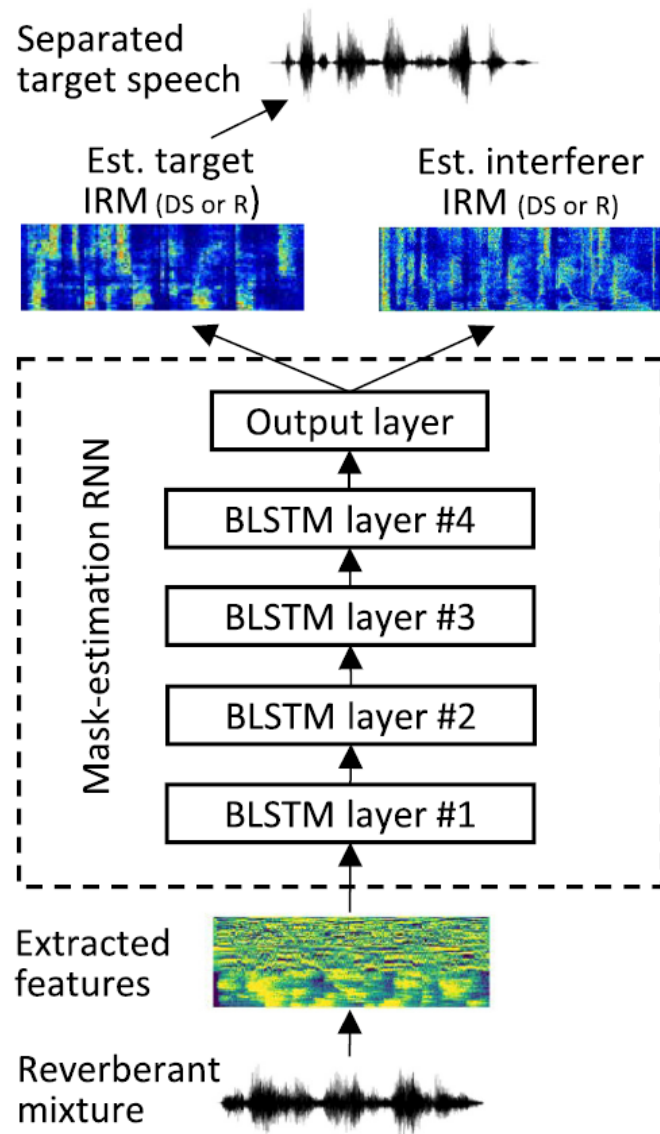
- Introduction
 - The cocktail party problem
 - Computational auditory scene analysis (CASA)
 - Speech separation as DNN based mask estimation
- **Talker-dependent speaker separation**
 - Human listener test
- **Talker-independent speaker separation**
 - Deep CASA approach

Speaker separation

- **Kinds of speaker separation**
 - Talker-dependent: Test speakers have been used during training
 - Target-dependent: Target speaker has been used during training, but not interfering speakers
 - Talker-independent: No test speaker has been used during training
- **Earlier work shows that DNN based mask estimation remains an effective approach to address talker or target dependent speaker separation (Huang et al.'14; Du et al.'14)**
- **We recently addressed reverberant talker-dependent speaker separation as DNN-based ideal ratio mask (IRM) estimation (Healy et al.'19)**
 - The IRM can be viewed as a soft version of the IBM

Ratio masking for reverberant speaker separation

- **We train a recurrent neural network (RNN) with bidirectional long short-term memory (BLSTM)**
 - To separate either direct-sound (DS) target speaker, or reverberant (R) target speaker
 - Target speaker at 1 m distance and interferer at 2 m, with $T60 = 0.6$ s



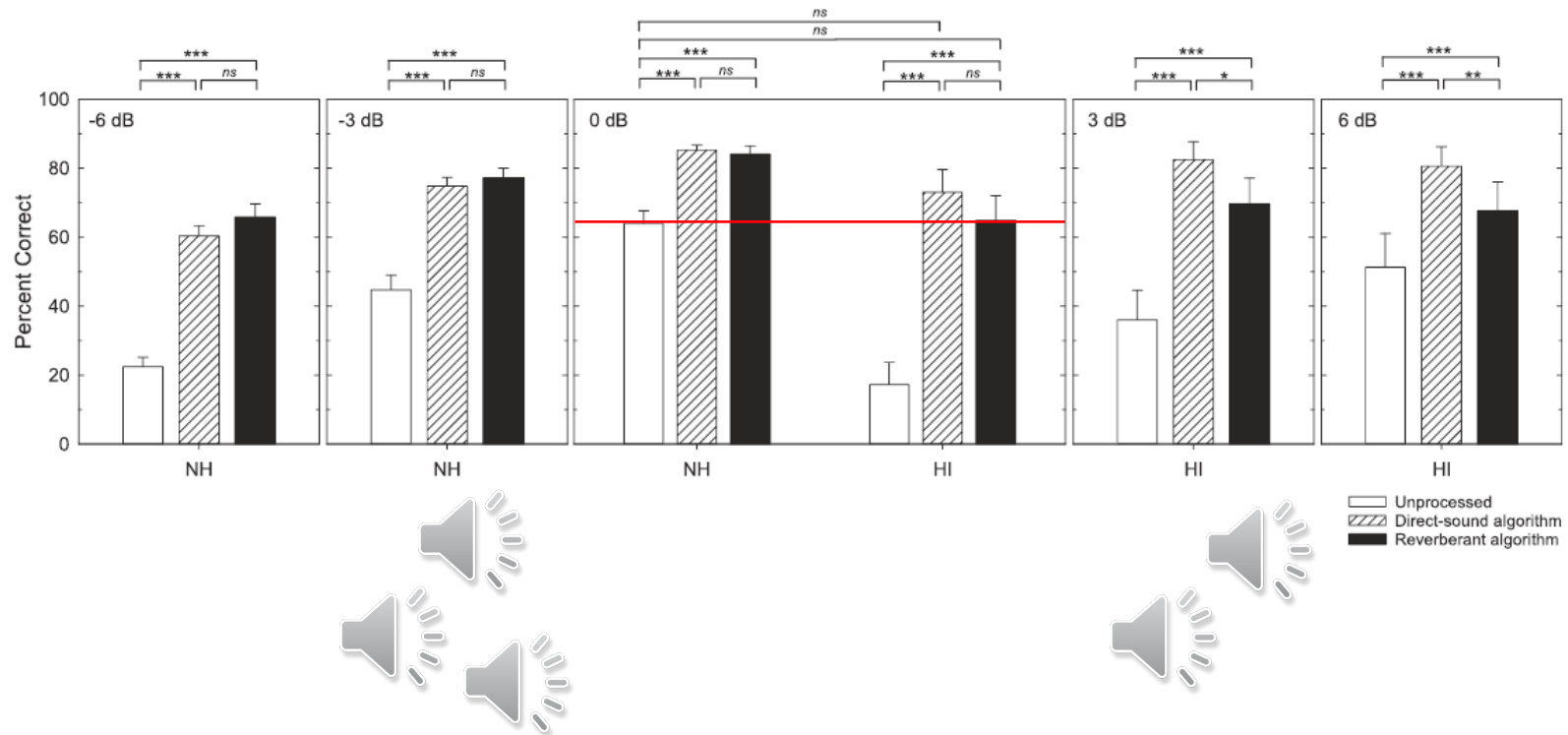
STOI and PESQ performance

- **For DS (anechoic) target speaker separation**

Input SNR (dB)	Unprocessed STOI	Processed STOI	Unprocessed PESQ	Processed PESQ
-6.00	45.6	78.7	1.54	2.30
-3.00	50.8	81.4	1.60	2.41
0.00	56.1	83.6	1.66	2.50
3.00	61.0	85.1	1.76	2.58
6.00	65.1	86.3	1.88	2.67

- Large STOI (standard intelligibility metric) and PESQ (standard speech quality metric) improvements are obtained by the speaker separation model

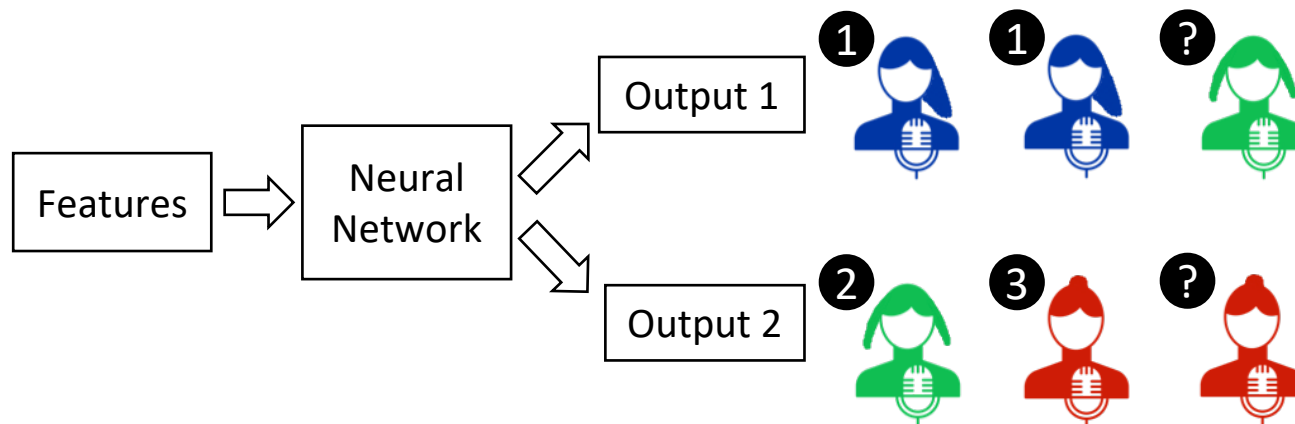
Listening test results and demos



- First demonstration of intelligibility improvements for HI listeners in both interfering speech and room reverberation
- At the common SNR of 0 dB, HI listeners with algorithm benefit outperform NH listeners without processing

Talker-independent speaker separation

- **This is the most general case, and it cannot be adequately addressed by training with many speaker pairs**
 - This is due to the permutation problem (Hershey et al.'16), referring to the ambiguity of output-layer and speaker correspondence during training



Deep clustering

- **Deep clustering (DC) is the first approach to talker-independent separation that combines DNN based supervised learning and clustering (Hershey et al.'16)**
- **With the ground truth partition of all T-F units, an affinity matrix is defined as**

$$A = YY^T$$

- Y is the indicator matrix built from the IBM. $Y_{i,c}$ is set to 1 if unit i belongs to (or dominated by) speaker c , and 0 otherwise
- $A_{i,j} = 1$ if units i and j belong to the same speaker, and 0 otherwise
- **To estimate the ground truth partition, DNN is trained to produce embedding vectors such that clustering in the embedding space provides a binary mask**

Permutation invariant training (PIT)

- **Recognizing that talker-dependent separation ties each DNN output to a specific speaker (permutation variant), PIT seeks to untie DNN outputs from speakers in order to achieve talker independence (Kolbak et al.'17)**
 - Specifically, for a pair of speakers, there are two possible assignments, each of which is associated with a mean squared error (MSE). The assignment with the lower MSE is chosen and the DNN is trained to minimize the corresponding MSE
- **Frame-level PIT (tPIT): Permutation can vary from frame to frame, hence needs speaker tracing (sequential grouping) for speaker separation**
- **Utterance-level PIT (uPIT): Permutation is fixed for a whole utterance, hence needs no speaker tracing**

Deep CASA approach

- **Limitations of deep clustering and PIT**
 - In deep clustering, embedding vectors for T-F units with similar energies from underlying speakers tend to be ambiguous
 - uPIT does not work as well as tPIT at the frame level, particularly for same-gender speakers, but tPIT requires speaker tracking
- **Speaker separation in CASA is talker-independent**
 - CASA typically performs simultaneous grouping first, and then sequential grouping across time
- **With these observations, we proposed a deep CASA approach (Liu & Wang'19)**
 - For simultaneous grouping, tPIT is employed to predict the spectra of underlying speakers at each frame
 - For sequential grouping, a sequence model is trained to assign simultaneously grouped spectra to underlying speakers

Simultaneous grouping

- **Input: complex short-time Fourier transform (STFT) of a mixture $Y(t, f)$**
- **Training target: complex ideal ratio mask (cIRM), in order to restore phase (Williamson et al.'16)**

$$X_c(t, f) = cIRM(t, f) \otimes Y(t, f)$$

- **Output: unorganized estimates of speakers $\hat{X}_c(t, f)$**

$$\hat{X}_c(t, f) = cRM(t, f) \otimes Y(t, f)$$

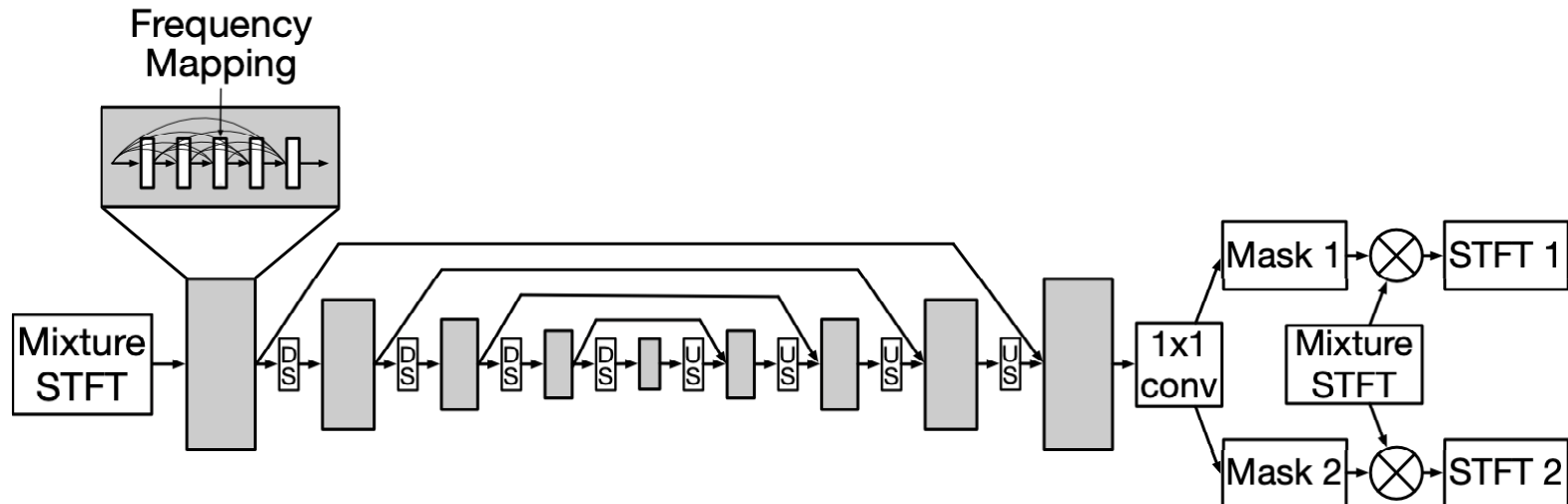
Simultaneous grouping (cont.)

- **Optimization objective: frame-level optimized tPIT training criterion, with time-domain SNR objective**
 - Organize $\hat{X}_c(t, f)$ with respect to the minimum frame-level loss, so that each organized output $\hat{X}_{\theta_c(t)}(t, f)$ corresponds to a single speaker
 - Apply inverse STFT to $\hat{X}_{\theta_c(t)}(t, f)$, and compute utterance-level SNR for the final time-domain estimates $\hat{x}_{\theta_c(t)}(n)$

$$J^{tPIT-SNR} = \sum_{c=1}^2 10 \log \frac{\sum_n x_c(n)^2}{\sum_n (x_c(n) - \hat{x}_{\theta_c(t)}(n))^2}$$

Simultaneous grouping network

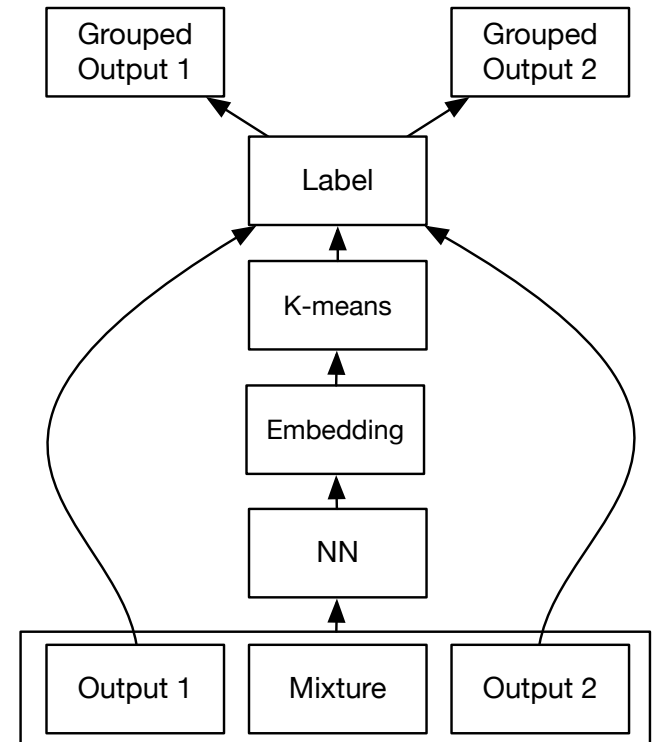
- **A DenseUNet is proposed for simultaneous grouping**
 - Downsampling (DS) layers, upsampling (US) layers and skip connections model global patterns and preserve fine-grained details
 - Densely connected convolutional layers exploit higher level of abstraction
 - Frequency mapping layers alleviate the inconsistency between different frequencies by reorganizing them in a new space



Sequential grouping

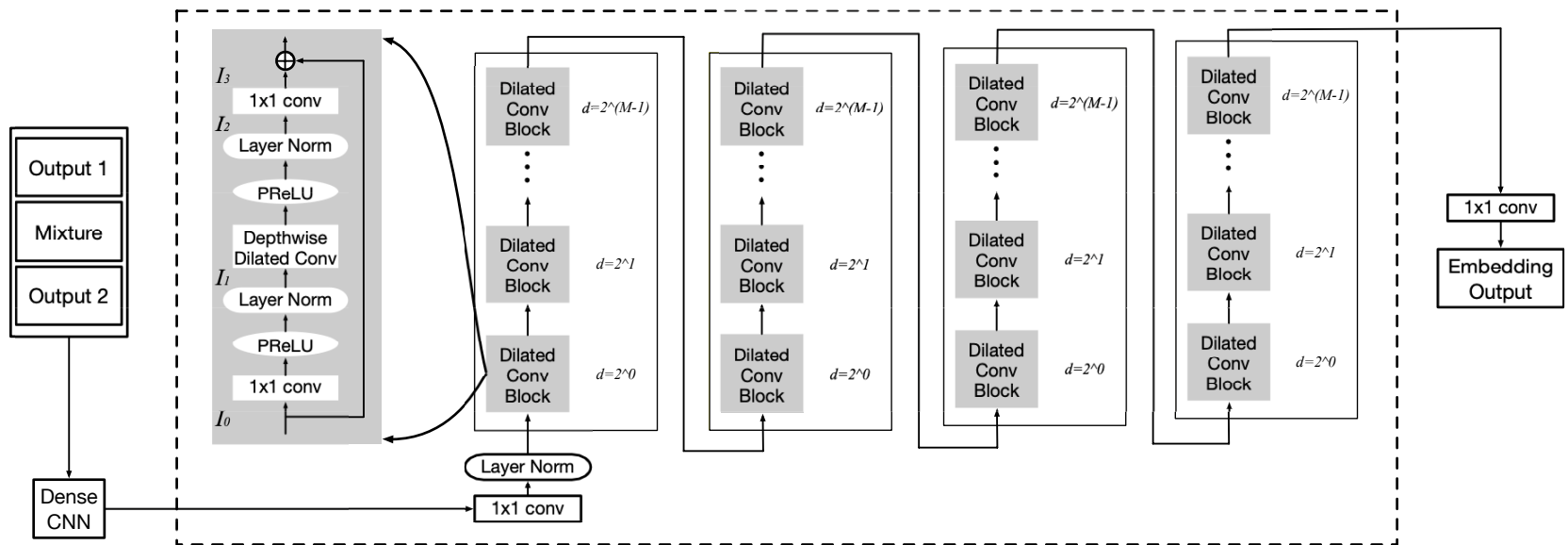
- **Input:** mixture spectrogram and two spectral estimates
- **Output:** frame-level embedding vectors, indicating output assignment $V(t)$
- **Indicator:** $A(t)$ is $[1 \ 0]$ if the minimum frame-level loss is achieved when tPIT Dense-UNet's output 1 is paired with Speaker 1, otherwise $A(t)$ is $[0 \ 1]$
- **Objective:**

$$J^{DC} = \|\mathbf{V}\mathbf{V}^T - \mathbf{A}\mathbf{A}^T\|_F^2$$



Sequential grouping network

- A temporal convolutional network (TCN) is employed as the sequence model
- Preprocessed features are fed to 8 consecutive dilated convolutional blocks, with an exponentially increasing dilation factor
- The 8 blocks are repeated 3 more times before embedding estimation



Evaluation

- **We use the WSJ0-2mix dataset, a monaural two-talker speaker separation dataset**
 - A 30-hour training set and a 10-hour validation set
 - A 5-hour open-condition (OC) test set, with untrained speakers
- **Models**
 - Deep CASA: tPIT Dense-UNet for simultaneous grouping and TCN for sequential grouping
 - uPIT Dense-UNet
- **Metrics: in addition to PESQ and extended STOI (ESTOI) we use**
 - Signal-to-distortion ratio improvement (Δ SDR)
 - Scale-invariant signal-to-noise ratio improvement (Δ SI-SNR)

Results

	Output Assign.	Δ SDR (dB)	PESQ	ESTOI (%)
Mixture	-	0.0	2.02	56.1
tPIT Dense-UNet	Default	0.0	1.99	55.8
	Optimal	19.1	3.63	94.3
uPIT Dense-UNet	Default	15.2	3.24	88.9
	Optimal	17.0	3.40	91.6

Deep
CASA



Sequential grouping comparison

Deep CASA	→	Simul. Group.	Seq. Group.	Frame Assign. Errors (%)
		tPIT Dense-UNet	TCN	1.38
		uPIT Dense-UNet	-	3.43

- **The results demonstrate the benefits of the proposed two-stage strategy, which optimizes frame-level separation and speaker tracking in turn, and achieves better performance in both objectives**

Speaker separation comparisons


- **Deep CASA produces very high quality speaker separation results, rivaling talker-dependent separation results**


	# of param.	Δ SDR (dB)	Δ SI-SNR (dB)	PESQ	ESTOI (%)
Mixture	-	0.0	0.0	2.02	56.1
uPIT [1]	92.7M	10.0	-	2.84	-
Conv-TasNet [2]	5.1M	15.6	15.3	3.24	-
Wang et al. [3]	56.6M	15.4	15.2	3.45	-
FurcaNeXt [4]	51.4M	18.4	-	-	-
Deep CASA	12.8M	18.0	17.7	3.51	93.2
IBM	-	13.8	13.4	3.28	89.1
IRM	-	13.0	12.7	3.68	92.9
PSM	-	16.7	16.4	3.98	96.0

[1] Kolbak et al.'17; [2] Luo & Mesgarani'19; [3] Wang et al.'19; [4] Shi et al.'19

Two-speaker separation demos


New pair of speakers 

Speaker1 - clean 

Speaker2 - clean 

Speaker1 -uPIT 









Speaker2 - uPIT 

Speaker1 – DC++ 

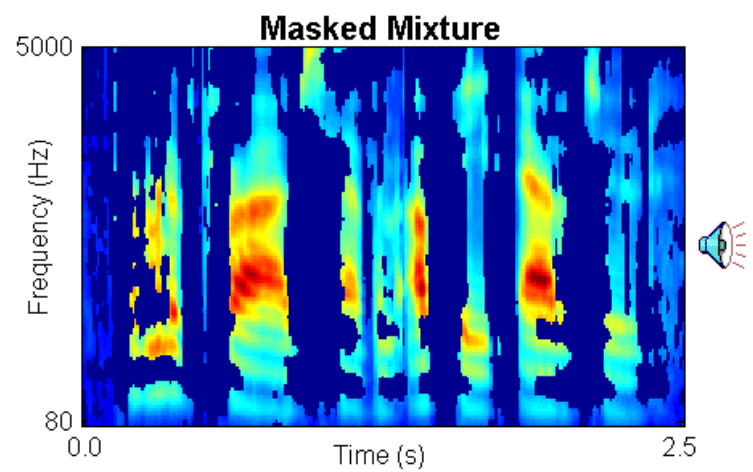
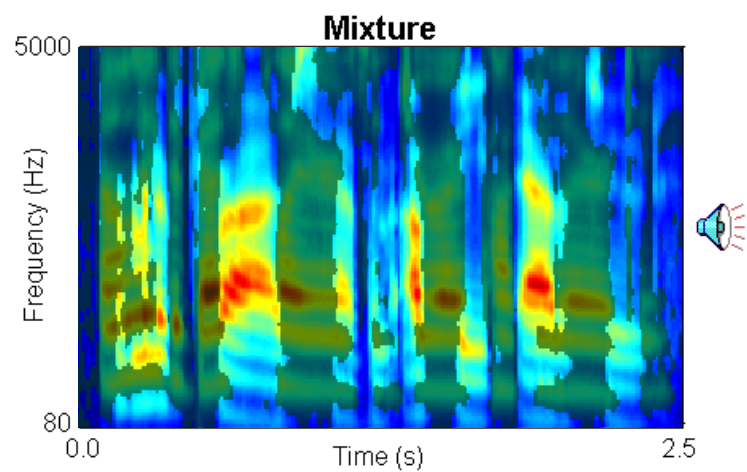
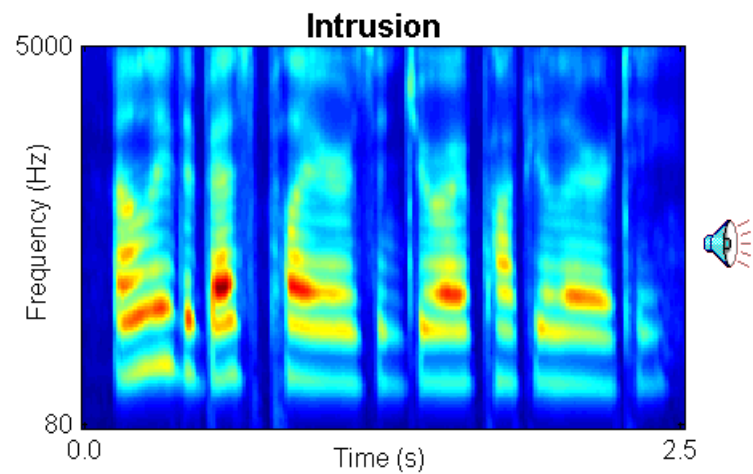
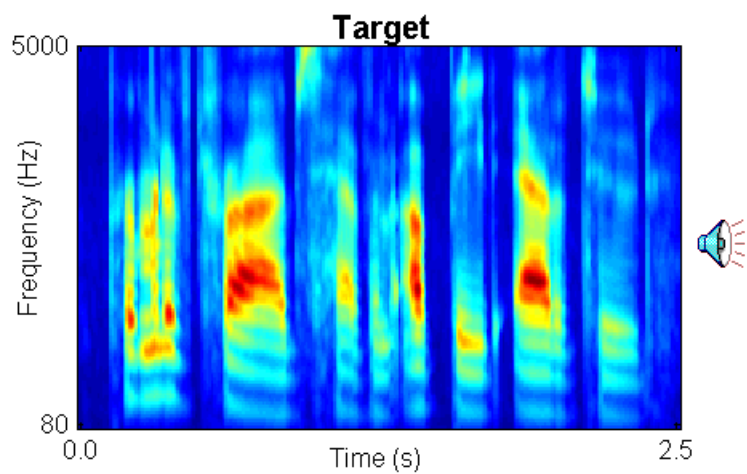
Speaker2 – DC++ 

Speaker1 – Deep CASA 

Speaker2 – Deep CASA 

Further demos	Mixture	Deep CASA	Clean
Noisy mixture (untrained)		 	 
Persian mixture (real recording)		 	

IBM illustration



Deep CASA

















Multi-speaker case

- **Extension to C concurrent speakers ($C > 2$) is straightforward in deep CASA**
 - Works well for speech mixtures with up to C speakers even without the prior knowledge about the speaker number

The following methods are trained and tested on WSJ0-3mix

	# of param.	Δ SDR (dB)	Δ SI-SNR (dB)	PESQ	ESTOI (%)
Mixture	-	0.0	0.0	1.66	38.5
uPIT	92.7M	7.7	-	-	-
Conv-TasNet	5.1M	13.1	12.7	2.61	-
Wang et al.	56.6M	12.5	12.1	2.77	-
Multi-speaker deep CASA	12.8M	14.6	14.3	2.77	80.8
IBM	-	13.6	13.3	2.86	82.1
IRM	-	13.0	12.6	3.44	88.6
PSM	-	16.8	16.4	3.80	93.7

Three-speaker separation demos

Description	Mixture	Deep CASA	Clean
Male-male-female (WSJ0-3mix OC)		  	  
Male-male-male (WSJ0-3mix OC)		  	  

A solution in sight for cocktail party problem?

- **What does a solution to the cocktail party problem look like?**
 - A system that achieves human auditory analysis performance in all listening situations (Wang & Brown'06)
- **An automatic speech recognition (ASR) system that matches the human speech recognition performance in all noisy environments**
 - Dependency on ASR

A solution in sight (cont.)?

- A speech separation system that helps hearing-impaired listeners to achieve the same level of speech intelligibility as normal-hearing listeners in *all noisy environments*
 - This is my current working definition – see my IEEE Spectrum cover story in March, 2017



Conclusion

- **Formulation of the cocktail party problem as mask estimation enables the use of supervised learning**
 - Supervised separation has yielded the first demonstrations of speech intelligibility improvement in noise and interfering speech in reverberation
- **A deep CASA framework for talker-independent speaker separation**
 - Simultaneous and sequential grouping mechanisms have been realized in a deep learning model
 - State-of-the-art performance in separating multi-speaker mixtures
 - Source code will be released soon
- **The cocktail party problem is within reach**