Towards Solving the Cocktail Party Problem

DeLiang Wang

Perception & Neurodynamics Lab
Ohio State University
& Northwestern Polytechnical University

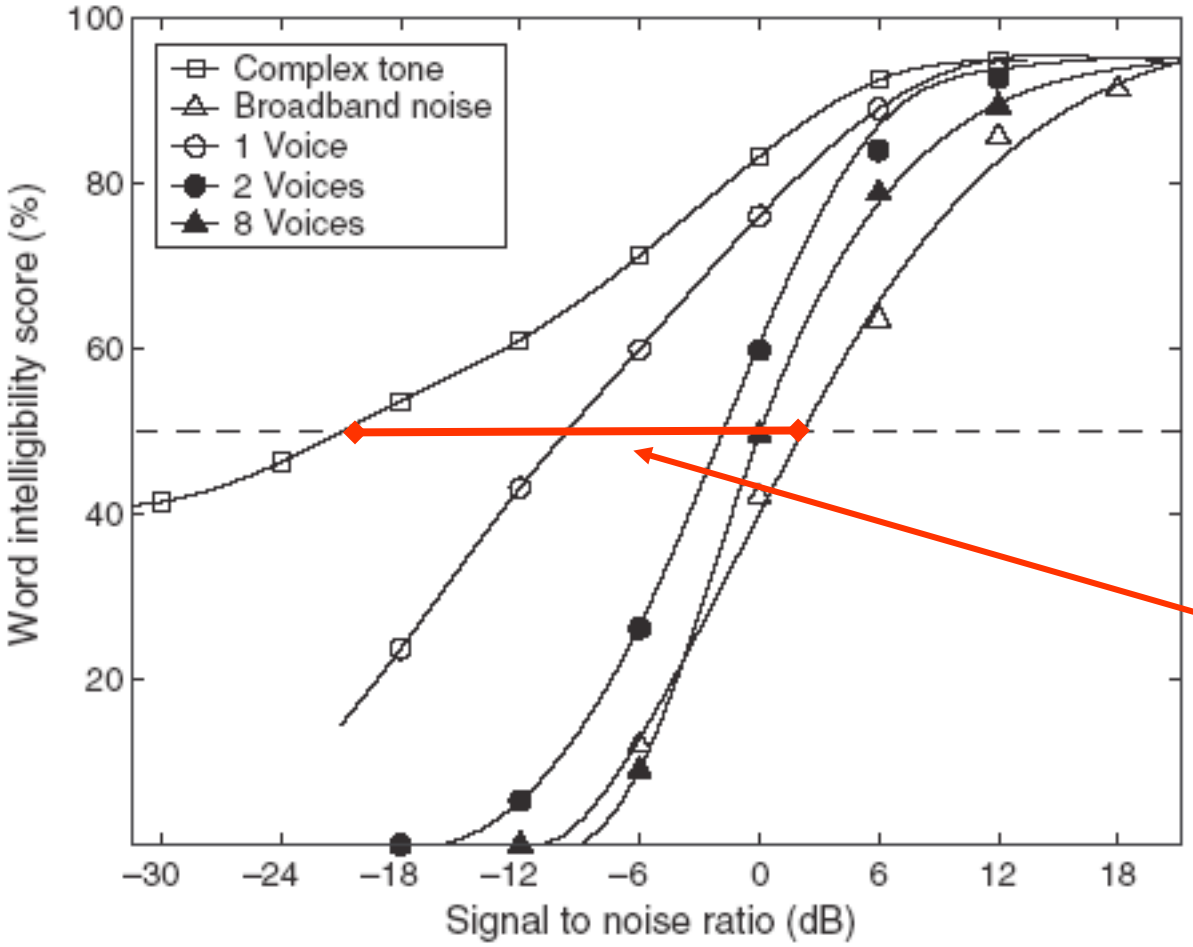
Outline of presentation

- **What is the cocktail party problem?**
  - Ideal binary mask and speech intelligibility
- **Speech separation as DNN based mask estimation**
  - Speech intelligibility tests on hearing impaired listeners
  - Generalization to new noises
    - Ideal ratio mask
- **Reverberant speech separation**
- **Speaker separation**
  - Talker-dependent
  - Talker-independent

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**

Human performance in different interferences



SRT 23 dB difference in Speech Reception Threshold!

Source: Wang & Brown (2006)

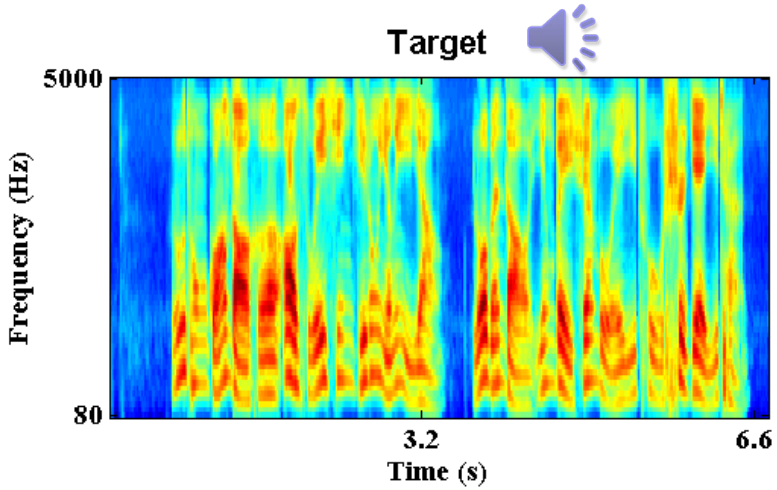
Ideal binary mask as a separation goal

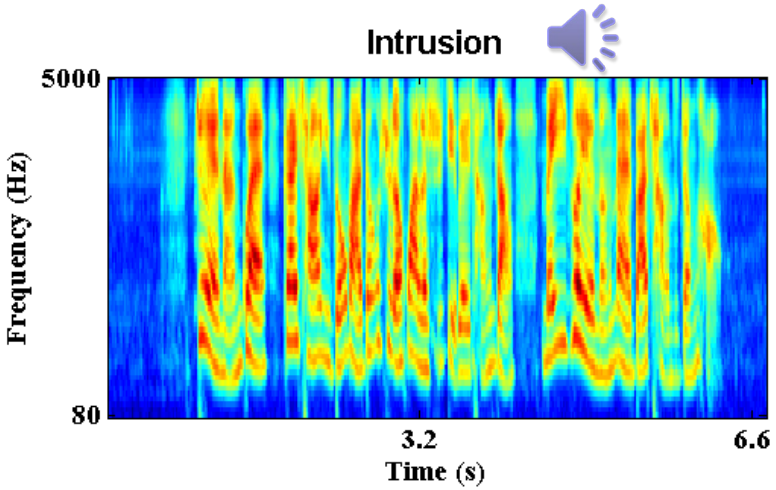
- **Motivated by the auditory masking phenomenon and auditory scene analysis, we suggested the ideal binary mask as a main goal of CASA (Hu & Wang, 2001; 2004)**
- **The idea is to retain parts of a mixture where the target sound is stronger than the acoustic background, and discard the rest**
- **The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

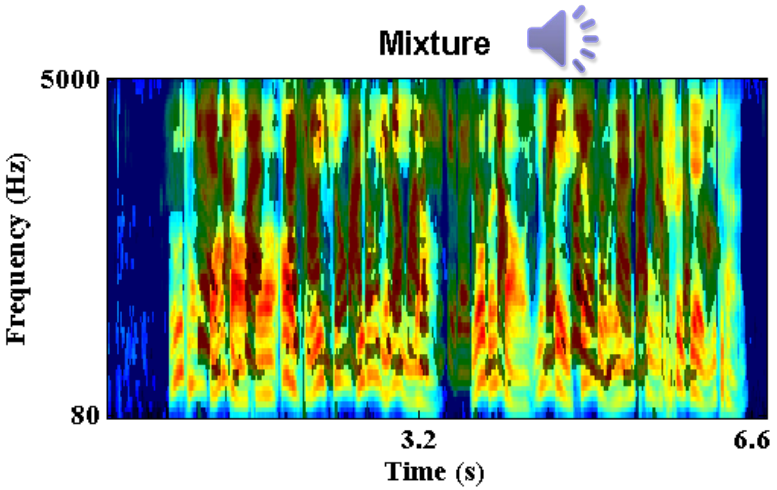
$$IBM(t,f) = \begin{cases} 1 & \text{if } SNR(t,f) \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

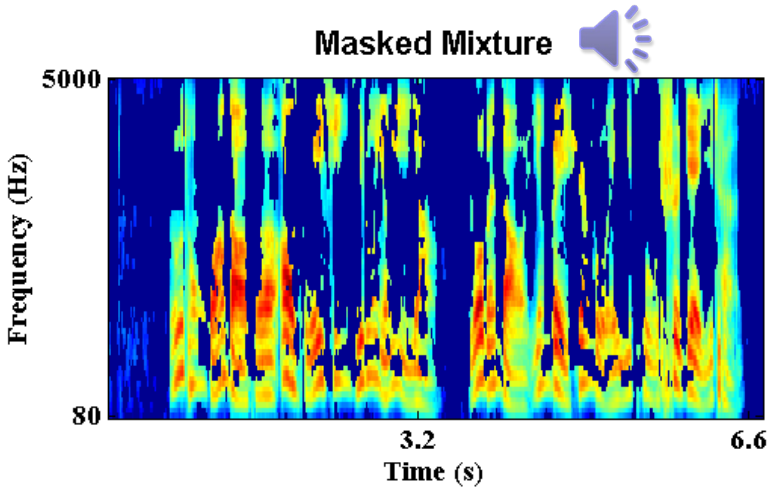
  - θ: A local SNR criterion (LC) in dB, which is typically chosen to be 0 dB
  - Optimal SNR: Under certain conditions the IBM with $\theta = 0$ dB is the optimal binary mask in terms of SNR gain (Li & Wang, 2009)
  - Maximal articulation index (AI) in a simplified version (Loizou & Kim, 2011)
  - It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

IBM illustration







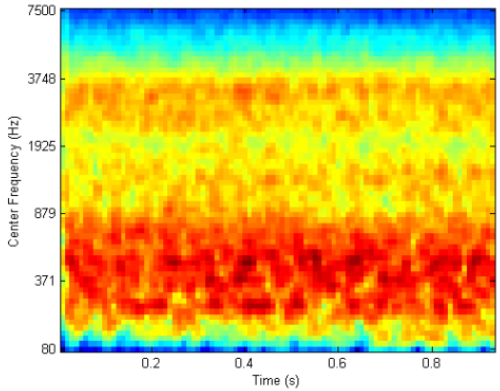


Subject tests of ideal binary masking

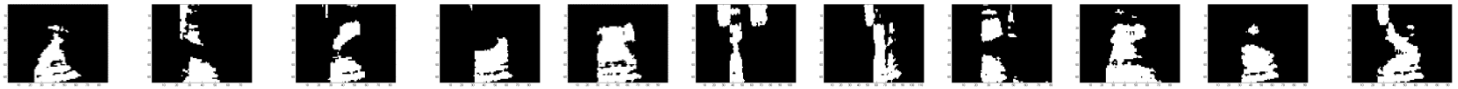
- **IBM separation leads to dramatic speech intelligibility improvements**
  - Improvement for stationary noise is above 7 dB for normal-hearing (NH) listeners (Brungart et al.'06; Li & Loizou'08; Ahmadi et al.'13; Chen'16), and above 9 dB for hearing-impaired (HI) listeners (Anzalone et al.'06; Wang 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 perception of noise with binary gains

- **Wang et al. (2008) found that, when LC is chosen to be the same as the input SNR, nearly perfect intelligibility is obtained when input SNR is -∞ dB (i.e. the mixture contains noise only with no target speech)**
- **IBM modulated noise for ???**



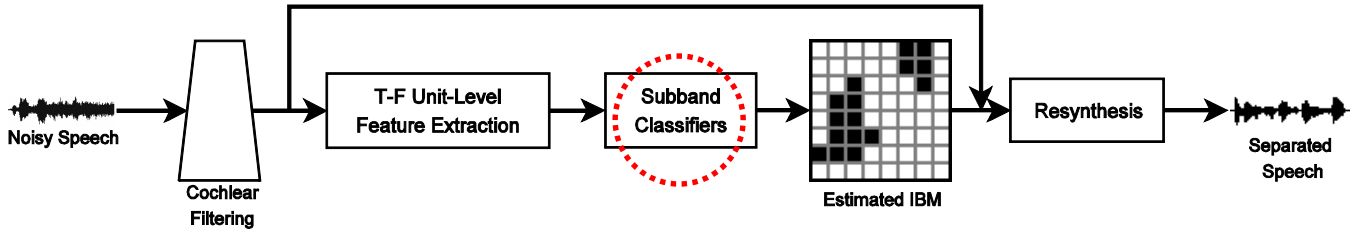
Speech shaped noise (SSN)



Outline of presentation

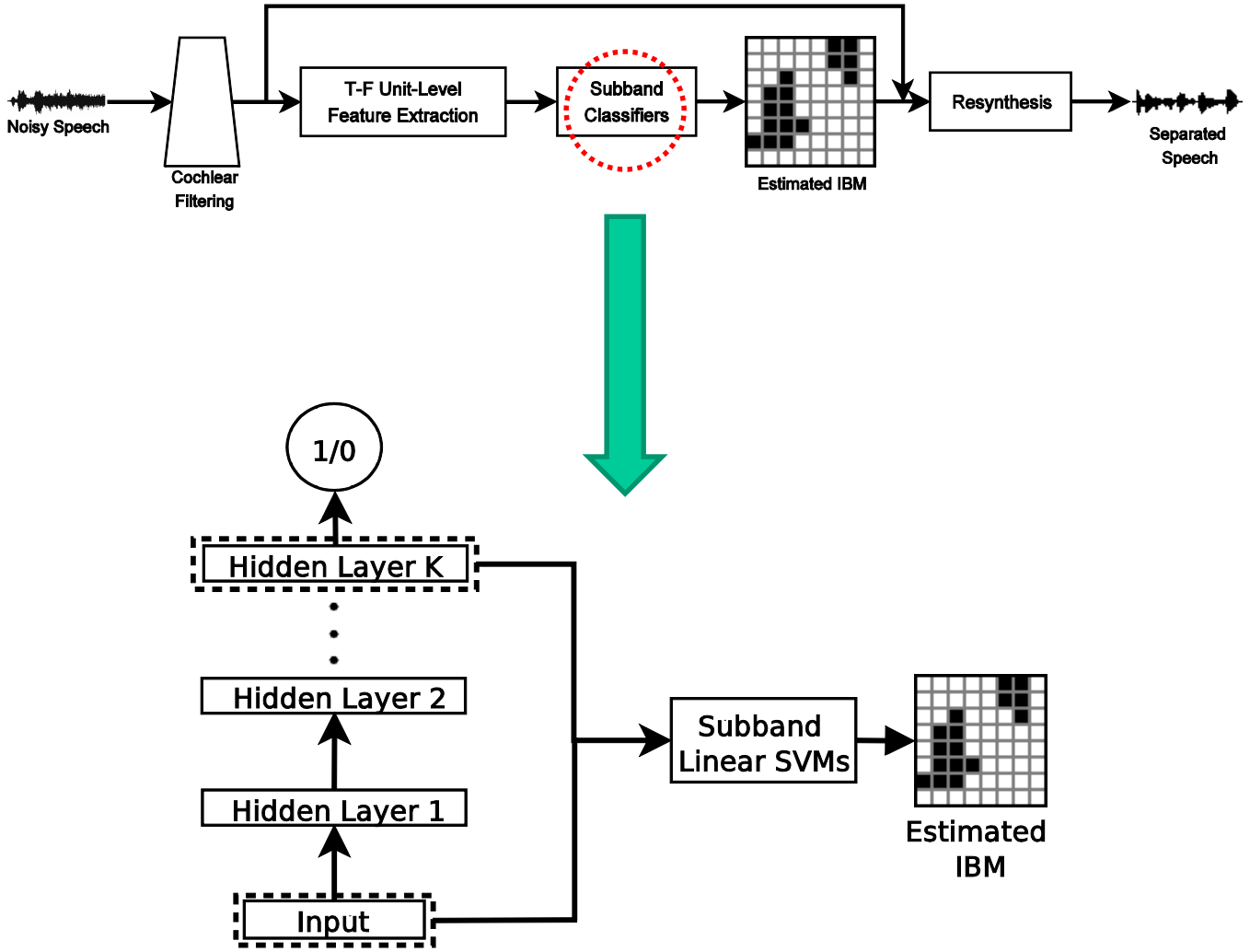
- **What is the cocktail party problem?**
  - Ideal binary mask and speech intelligibility
- **Speech separation as DNN based mask estimation**
  - Speech intelligibility tests on hearing impaired listeners
  - Generalization to new noises
    - Ideal ratio mask
- **Reverberant speech separation**
- **Speaker separation**
  - Talker-dependent
  - Talker-independent

DNN as subband classifier (Wang & Wang'13)



- **Y. Wang and Wang (2013) first introduced DNN to address the speech separation problem**
  - DNN is used for as a subband classifier, performing feature learning from raw acoustic features
  - Classification aims to estimate the IBM

DNN as subband classifier (Wang & Wang'13)

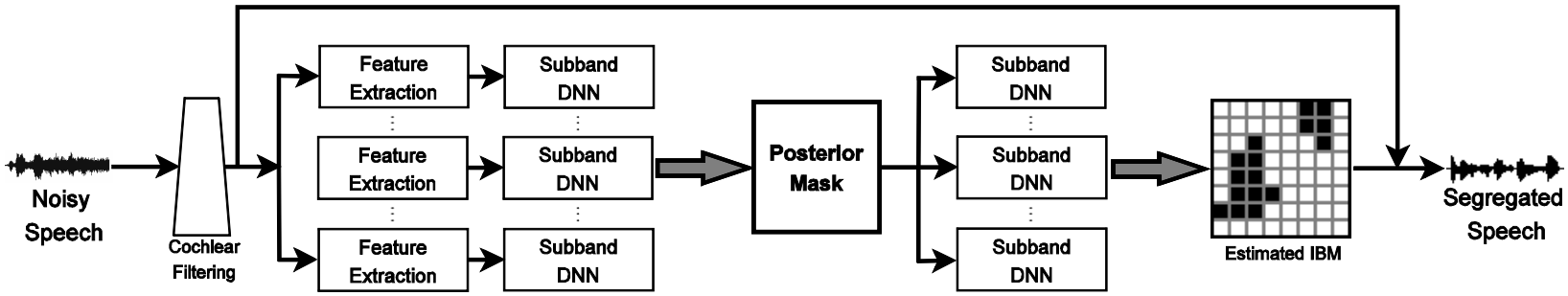


Extensive training with DNN

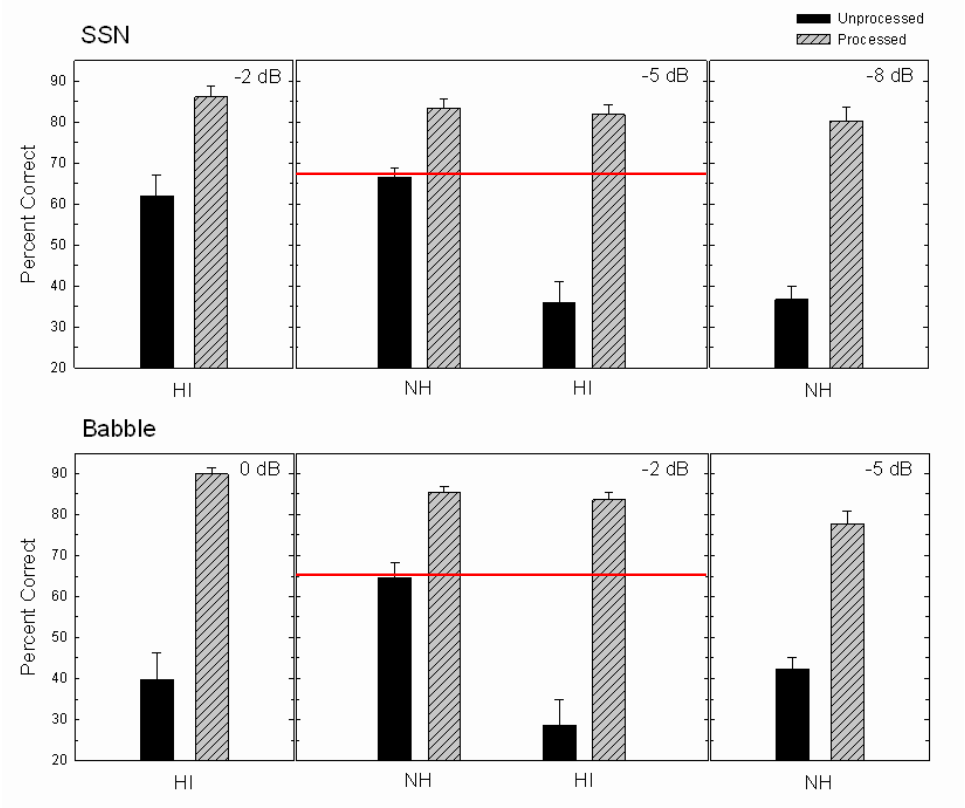
- **Training on 200 randomly chosen utterances from both male and female IEEE speakers, mixed with 100 environmental noises at 0 dB (~17 hours long)**
  - Six million fully dense training samples in each channel, with 64 channels in total
- **Evaluated on 20 unseen speakers mixed with 20 unseen noises at 0 dB**
- **DNN based classifier produced the state-of-the-art separation results at the time**

Speech intelligibility evaluation

- **Healy et al. (2013) subsequently evaluated the classifier on speech intelligibility of hearing-impaired listeners**
  - A very challenging problem: "The interfering effect of background noise is the single greatest problem reported by hearing aid wearers" (Dillon'12)
- **Two stage DNN training to incorporate T-F context in classification**



Results and sound demos



- **Both HI and NH listeners showed intelligibility improvements**
- **HI subjects with separation outperformed NH subjects without separation**

Generalization to new noises

- **While previous speech intelligibility results are impressive, a major limitation is that training and test noise samples were drawn from the same noise segments**
  - Speech utterances were different
  - Noise samples were randomized
- **This limitation can be addressed through large-scale training for IRM estimation (Chen et al.'16)**

$$IRM(t,f) = \sqrt{\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}} = \sqrt{\frac{SNR(t,f)}{SNR(t,f)+1}}$$

  - IRM can be viewed as a soft version of the IBM

Large-scale training

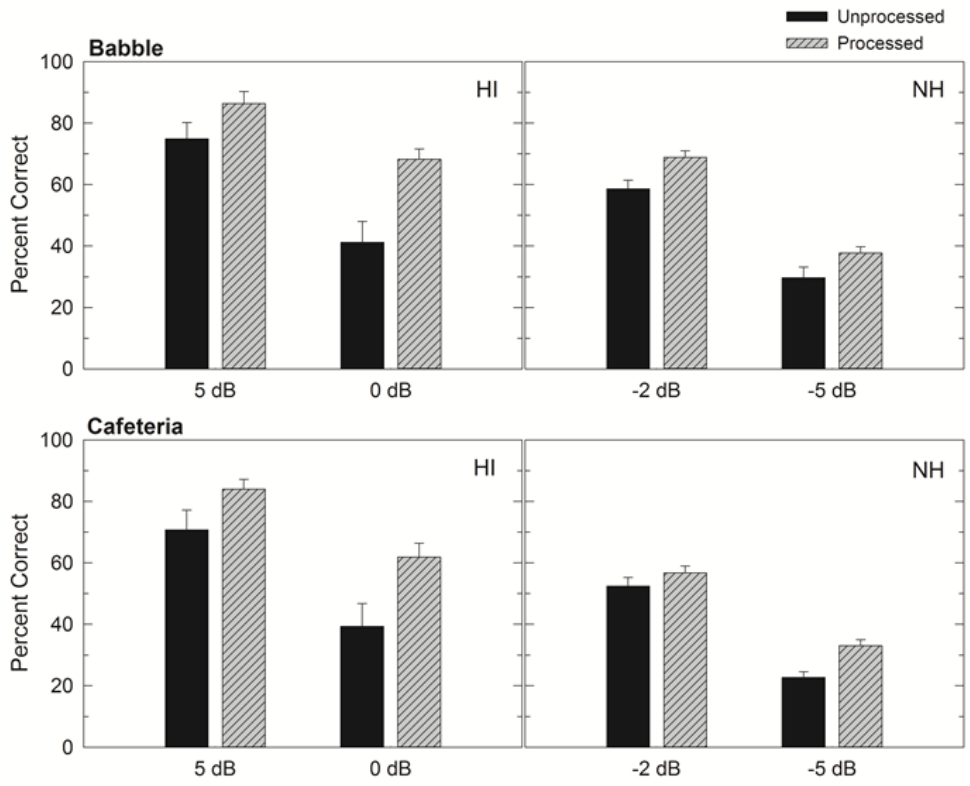
- **Training set consisted of 560 IEEE sentences mixed with 10,000 (10K) non-speech noises (a total of 640,000 mixtures)**
  - The total duration of the noises is about 125 h, and the total duration of training mixtures is about 380 h
  - Training SNR is fixed to -2 dB
- **The only feature used is the simple T-F unit energy**
- **DNN architecture consists of 5 hidden layers, each with 2048 units**
- **Test utterances and noises are both different from those used in training**

STOI performance at -2 dB input SNR

	Babble	Cafeteria	Factory	Babble2	Average
Unprocessed	0.612	0.596	0.611	0.611	0.608
100-noise model	0.683	0.704	0.750	0.688	0.706
10K-noise model	0.792	0.783	0.807	0.786	0.792
Noise-dependent model	0.833	0.770	0.802	0.762	0.792

- **STOI is a standard metric for predicted speech intelligibility**
- **DNN model with large-scale training provides similar results to noise-dependent model**

Listening test results



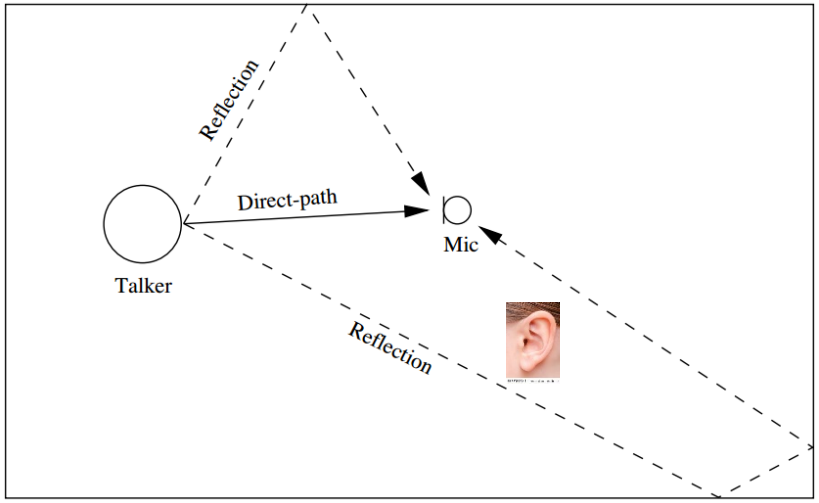
- Both NH and HI listeners received benefit from algorithm processing in all conditions, with larger benefits for HI listeners

Outline of presentation

- **What is the cocktail party problem?**
  - Ideal binary mask and speech intelligibility
- **Speech separation as DNN based mask estimation**
  - Speech intelligibility tests on hearing impaired listeners
  - Generalization to new noises
    - Ideal ratio mask
- **Reverberant speech separation**
- **Speaker separation**
  - Talker-dependent
  - Talker-independent

Reverberation

- **Reverberation is everywhere: Reflections of the sound source (direct sound) from various surfaces**



- **Reverberation and background noise have confounding effects and can severely degrade speech intelligibility for HI listeners (Nabelek & Mason'81)**

Signal model

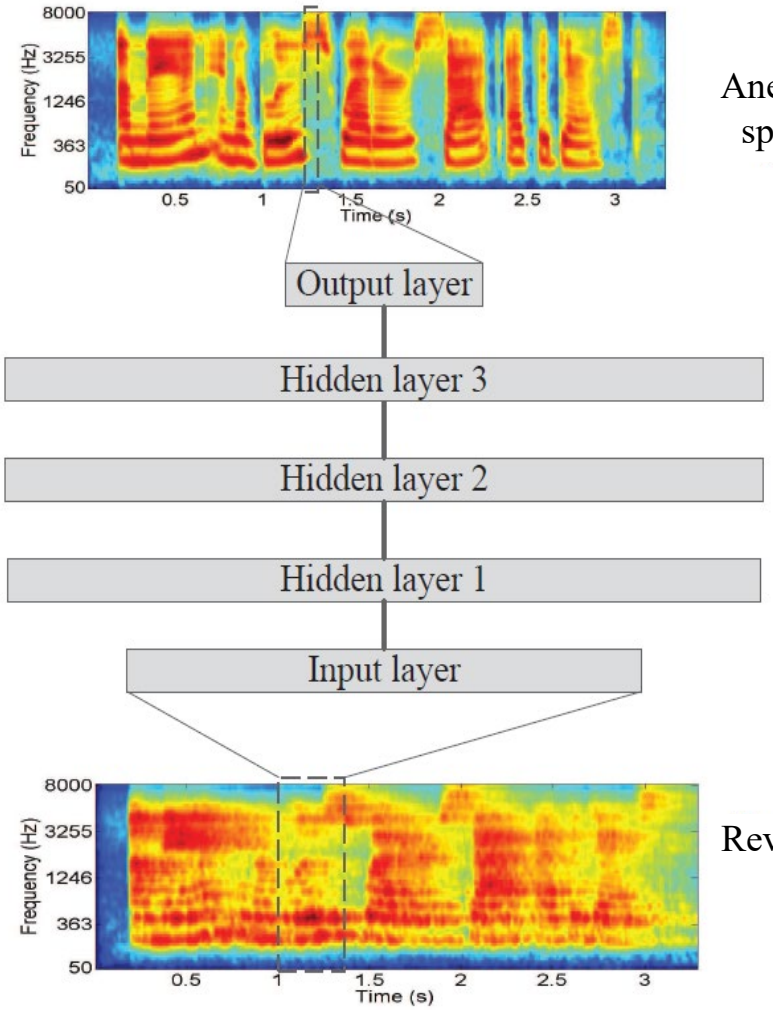
- **Reverberant-noisy signal model**

$$y(t) = s(t) * h(t) + n(t)$$

  - Clean (anechoic) speech: $s(t)$
  - Room impulse response: $h(t)$
  - Background noise: $n(t)$
  - Reverberant-noisy speech signal: $y(t)$

Previous spectral mapping work

- **Han et al. (2015) conducted the first DNN study on dereverberation and denoising**
  - DNN was trained to learn the mapping from the spectrum of reverberant-noisy speech to the spectrum of anechoic speech (Lu et al.'13; Xu et al.'14; Han et al.'14)
  - However, an informal listening test indicates no speech intelligibility gain for HI listeners



Enhancement of reverberant-noisy speech as mask estimation

- **Zhao et al. (2018) recently approached dereverberation and denoising as IRM estimation**
- **The input is a set of complementary features**
- **DNN architecture**
  - Feedforward network with four hidden layers, each with 2048 units
- **Speech intelligibility evaluation of enhanced speech on HI and NH listeners**
  - At reverberation time of 0.6 s

Sound demos - typical stimuli

- SSN (-5 dB)

Example 1	
unprocessed	
processed	
clean	

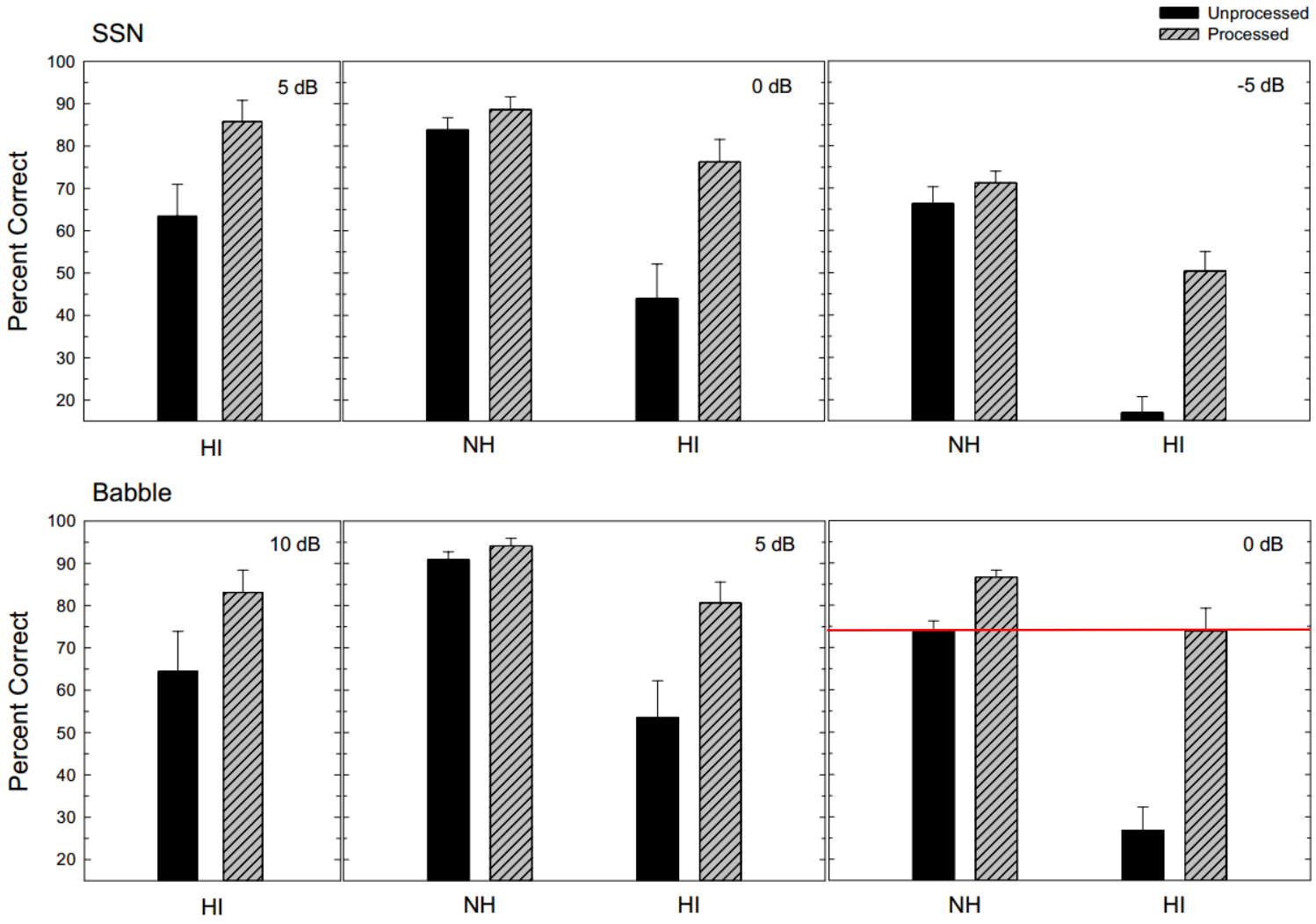
Example 2	
unprocessed	
processed	
clean	

- Babble (0 dB)

Example 3	
unprocessed	
processed	
clean	

Example 4	
unprocessed	
processed	
clean	

Intelligibility results



Outline of presentation

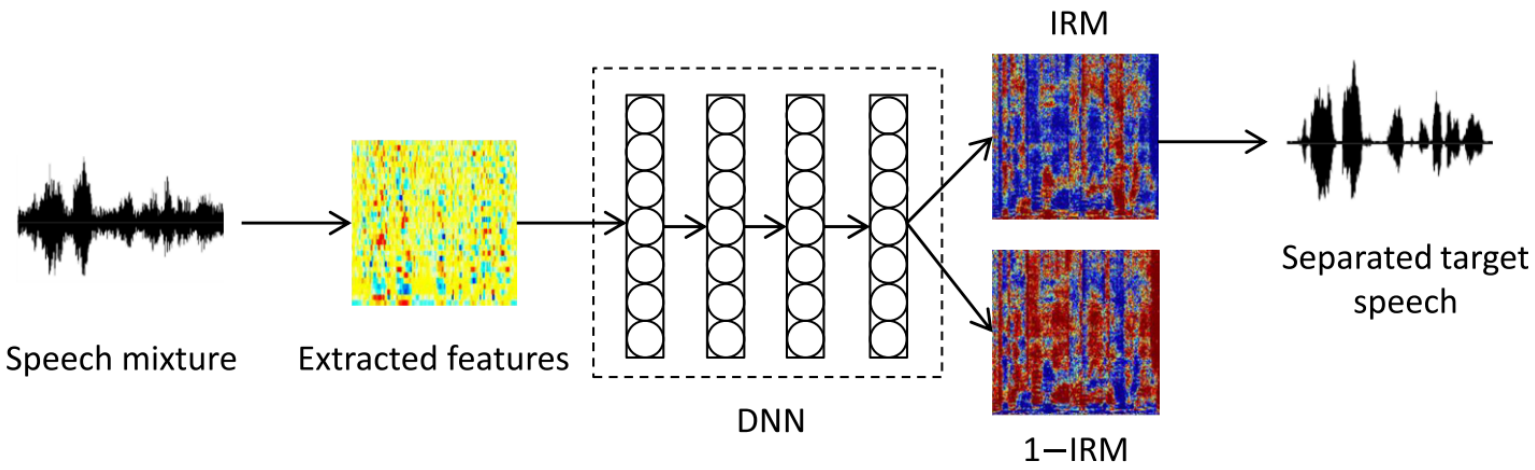
- **What is the cocktail party problem?**
  - Ideal binary mask and speech intelligibility
- **Speech separation as DNN based mask estimation**
  - Speech intelligibility tests on hearing impaired listeners
  - Generalization to new noises
    - Ideal ratio mask
- **Reverberant speech separation**
- **Speaker separation**
  - Talker-dependent
  - Talker-independent

Speaker separation

- **Unlike speech enhancement, speaker separation aims to extract the target speaker from a multi-talker mixture**
- **Earlier work shows that DNN based IBM/IRM estimation remains an effective approach to address speaker separation (Huang et al.'14; Du et al.'14)**
- **We recently addressed two-talker separation as DNN-based IRM estimation (Healy et al.'17)**

IRM estimation for speaker separation

- **In the case of two-talker mixtures, the complement of a ratio mask gives an estimate of the interfering talker**

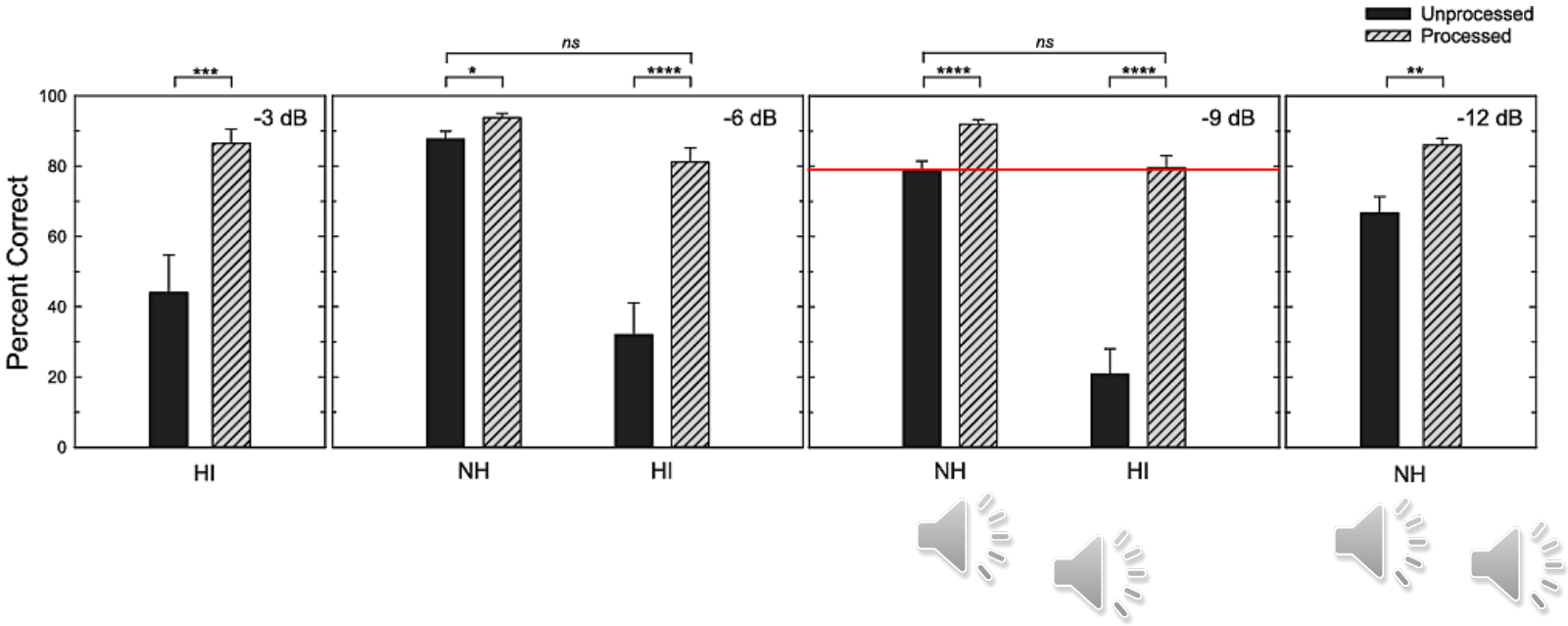


STOI performance

Input SNR (dB)	Output SNR (dB)	Unprocessed STOI	Processed STOI
-12.00	5.10	43.5	83.5
-9.00	6.09	49.9	86.6
-6.00	7.11	56.9	89.1
-3.00	8.20	64.2	91.1

- **Large STOI improvements are obtained by the speaker separation model**

Listening test results and demos



- First demonstration of intelligibility improvements for HI listeners in interfering speech
- At the common SNR of -6 and -9 dB, HI listeners with algorithm benefit perform equally well with 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**
- **Talker-independent separation can be treated as unsupervised clustering (Bach & Jordan'06; Hu & Wang'13)**
  - Such clustering, however, does not benefit from discriminant information utilized in supervised training
- **Deep clustering (Hershey et al.'16) is the first approach to talker-independent separation by combining DNN based supervised feature learning and spectral clustering**

Deep clustering

- **With the ground truth partition of all T-F units, an affinity matrix is defined as**

$$\boldsymbol{A = YY^T}$$

  - $\boldsymbol{Y}$ 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 better partition estimate**
- **Isik et al. (2016) extend deep clustering by incorporating an enhancement network after binary mask estimation, and performing end-to-end training of embedding and clustering**

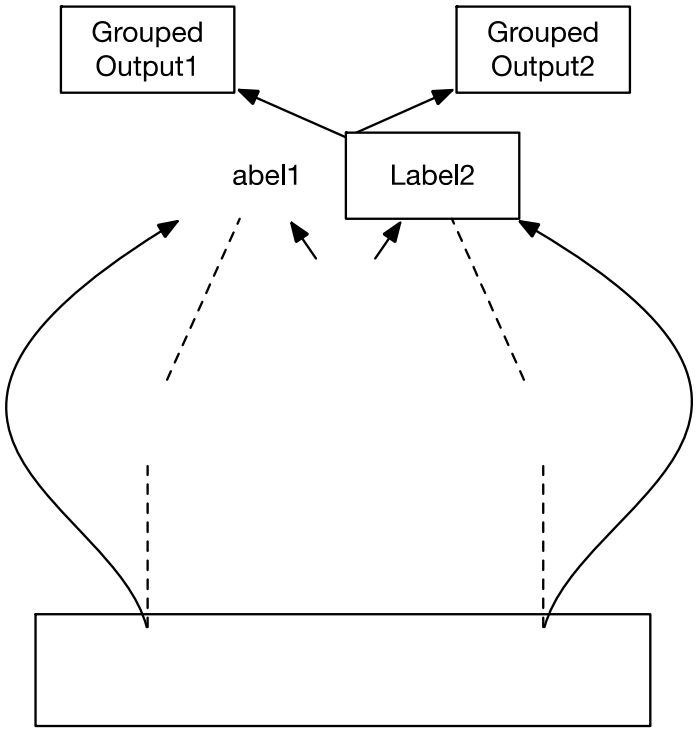
Permutation invariant training (PIT)

- **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
- **Two versions of PIT**
  - 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

CASA based 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 tracing
- **Speaker separation in CASA is talker-independent**
  - CASA performs simultaneous (spectral) grouping first, and then sequential grouping across time
- **Liu & Wang (2018) proposed a CASA based approach by leveraging PIT and deep clustering**
  - For simultaneous grouping, tPIT is trained to predict the spectra of underlying speakers at each frame
  - For sequential grouping, DNN is trained to predict embedding vectors for simultaneously grouped spectra

Sequential grouping in CASA



- **Differences from deep clustering**
  - In deep clustering, DNN based embedding is done at the T-F unit level, whereas it is done at the frame level in CASA
  - Constrained K-means in CASA ensures that the simultaneously separated spectra of the same frame are assigned to different speakers

Speaker separation performance

- **Talker-independent separation produces high-quality speaker separation results, rivaling talker-dependent separation results**

SDR improvements (in dB)

	Same Gender	Different Gender	Overall
Deep clustering++ (Isik et al., 2016)	9.4	12.0	10.8
PIT (Kolbæk et al., 2017)	7.5	12.2	10.0
CASA (Liu and Wang, 2018)	10.3	12.6	11.5

Talker-independent speaker separation demos

New pair of male-male speaker mixture 

Speaker1 -uPIT		Speaker2 - uPIT	
Speaker1 – DC++		Speaker2 – DC++	
Speaker1 - CASA		Speaker2 - CASA	
Speaker1 - clean		Speaker2 - clean	

Rapid advances in talker-independent speaker separation

ΔSI-SDR (dB), ΔSDR (dB) and PESQ comparison between recent algorithms on the open speaker condition of wsj0-2mix and wsj0-3mix. ("-" means not reported)

Approaches	wsj0-2mix			wsj0-3mix		
	ΔSI-SDR	ΔSDR	PESQ	ΔSI-SDR	ΔSDR	PESQ
Unprocessed	0.0	0.0	2.01	0.0	0.0	1.66
DC++ (Isik *et al*., 2016)	10.8	-	-	7.1	-	-
uPIT-ST (Kolbaek *et al*., 2017)	-	10.0	-	-	7.7	-
ADANet (Luo *et al*. 2018a)	10.4	10.8	2.82	9.1	9.4	2.16
Chimera++ (BLSTM) (Wang *et al*., 2018a)	11.2	11.5	-	-	-	-
+WA-MISI-5 (Wang *et al*., 2018b)	12.6	12.9	-	-	-	-
+ Filterbank learning (Wichern *et al*., 2018)	12.8	-	-	-	-	-
+ PhaseBook (Le Roux *et al*., 2018)	12.6	-	-	-	-	-
BLSTM-TasNet (Luo *et al*., 2018b)	13.2	13.6	3.04	-	-	-
conv-TasNet-gLN (Luo *et al*., 2018c)	14.6	15.0	3.25	11.6	12.0	2.50
Sign Prediction Net (Wang et *al*., 2018c)	**15.3**	**15.6**	**3.36**	**12.1**	**12.5**	**2.64**

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
  - Large-scale training with DNN is a promising direction to make speech separation perform in a variety of conditions
- **Reverberant-noisy speech separation can be effectively addressed in the same framework**
- **Major advances in both talker-dependent and talker-independent speaker separation**
- **The cocktail party problem is within reach**