From Auditory Masking to Binary Classification: Machine Learning for Speech Separation

DeLiang Wang

Perception & Neurodynamics Lab

Ohio State University

Outline of presentation

- **Auditory masking and speech intelligibility**
  - Ideal binary mask
  - Separation as binary classification
- **GMM based classification**
  - Speech intelligibility tests on normal hearing listeners
- **DNN based classification**
  - Speech intelligibility tests on hearing impaired listeners
- **Discussion: Problems of SNR**

What is auditory masking?

- **Definition: “The process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound” (American Standards Association, 1960)**
    - A basic phenomenon in auditory perception
- **Our daily experience that a sound is rendered inaudible or suppressed by its acoustic background**
    - In a way, separation is about unmasking or release from masking

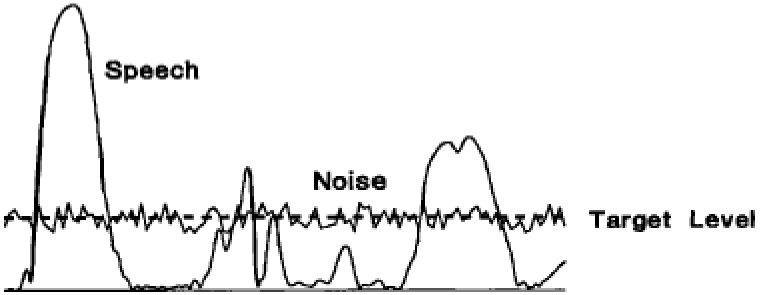
Masking and critical band

- **Fletcher (1940) introduced the concept of critical bands to describe the masking of a pure tone by wideband noise, leading to auditory filters**

- **Roughly speaking, within a critical band a stronger signal masks a weaker one**

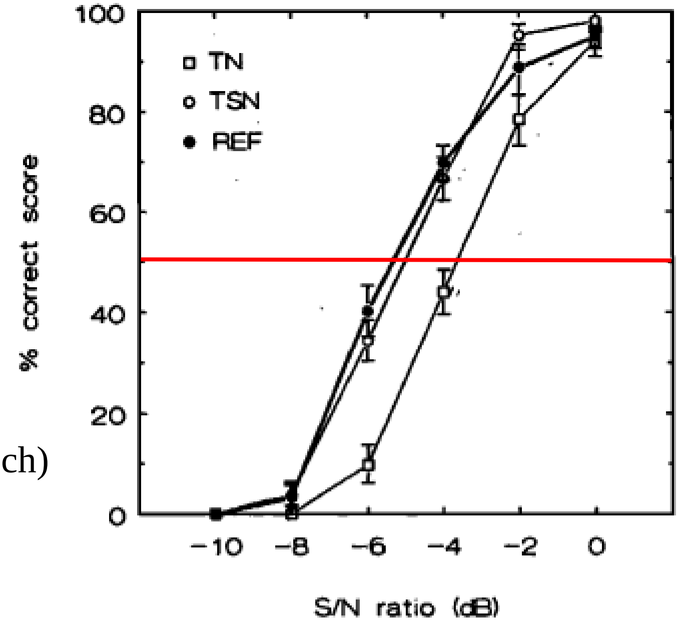
Masking and speech intelligibility

- **Drullman (1995) studied the intelligibility effects of speech below and above noise level**



Auditory filter output

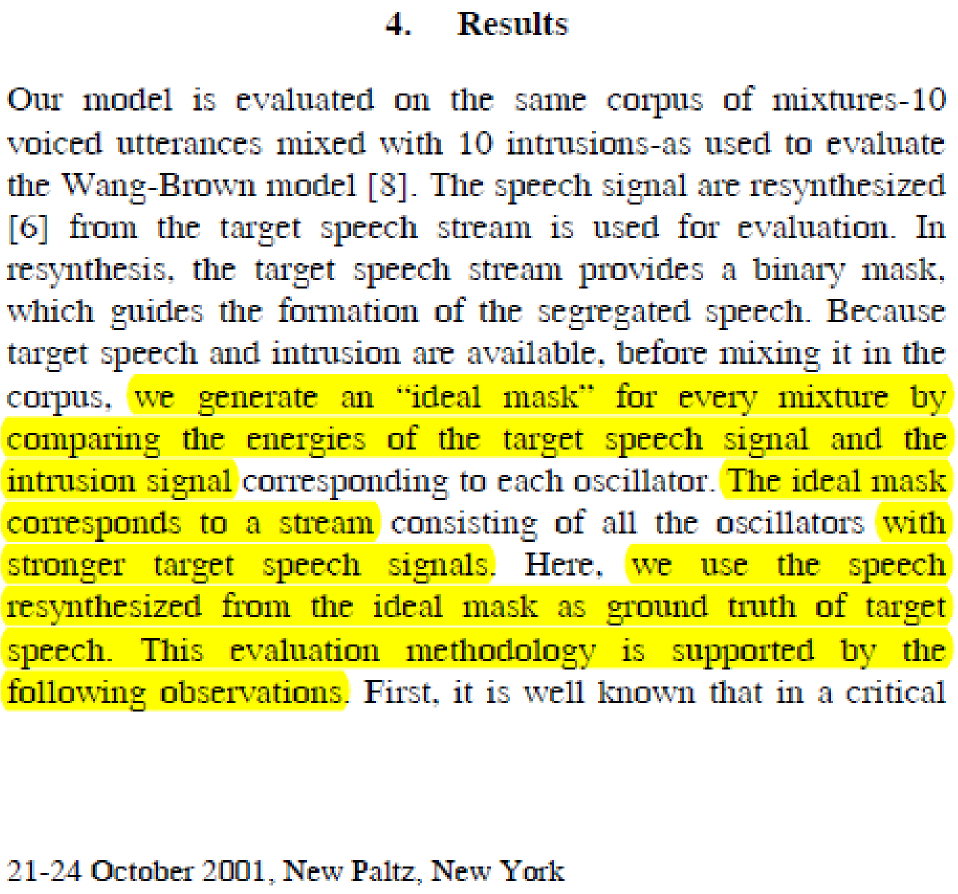
TN: Troughs with Noise (peaks with speech)
TSN: Troughs with Speech & Noise (peaks with speech)
REF: Original Mixture



- **Main findings**
  - Removing noise underneath speech has no benefit
  - Removing speech underneath noise has 2 dB detriment

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'01, WASPAA)**



4. Results

Our model is evaluated on the same corpus of mixtures-10 voiced utterances mixed with 10 intrusions-as used to evaluate the Wang-Brown model [8]. The speech signal are resynthesized [6] from the target speech stream is used for evaluation. In resynthesis, the target speech stream provides a binary mask, which guides the formation of the segregated speech. Because target speech and intrusion are available, before mixing it in the corpus, we generate an "ideal mask" for every mixture by comparing the energies of the target speech signal and the intrusion signal corresponding to each oscillator. The ideal mask corresponds to a stream consisting of all the oscillators with stronger target speech signals. Here, we use the speech resynthesized from the ideal mask as ground truth of target speech. This evaluation methodology is supported by the following observations. First, it is well known that in a critical

21-24 October 2001, New Paltz, New York

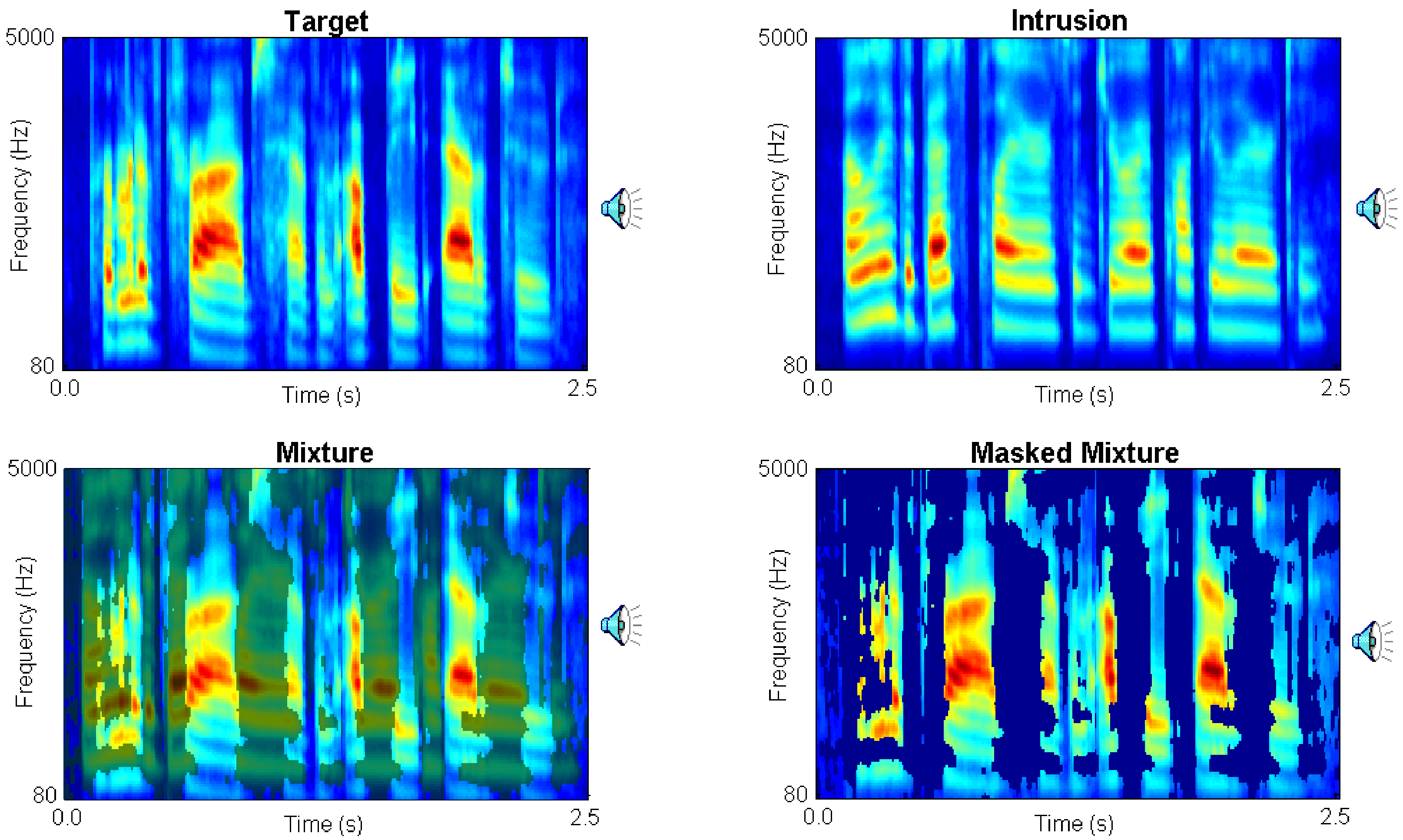
IBM definition

- **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 θ = 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



Target

Intrusion

Mixture

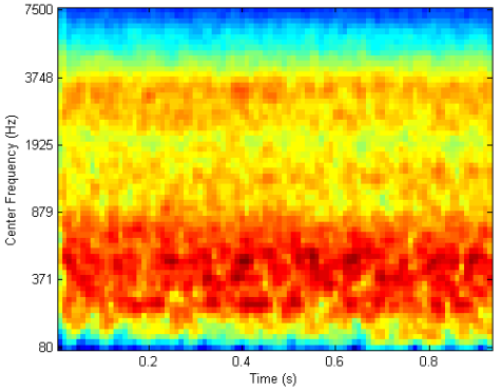
Masked Mixture

Subject tests of ideal binary masking

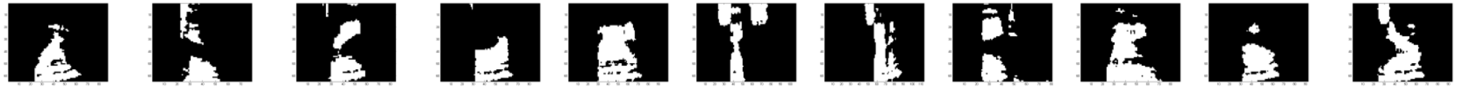
- **IBM separation leads to large speech intelligibility improvements**
  - Improvement for stationary noise is above 7 dB for normal-hearing (NH) listeners (Brungart et al.’06; Li & Loizou’08; Cao et al.’11; Ahmadi et al.’13), and above 9 dB for hearing-impaired (HI) listeners (Anzalone et al.’06; Wang 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 ???**



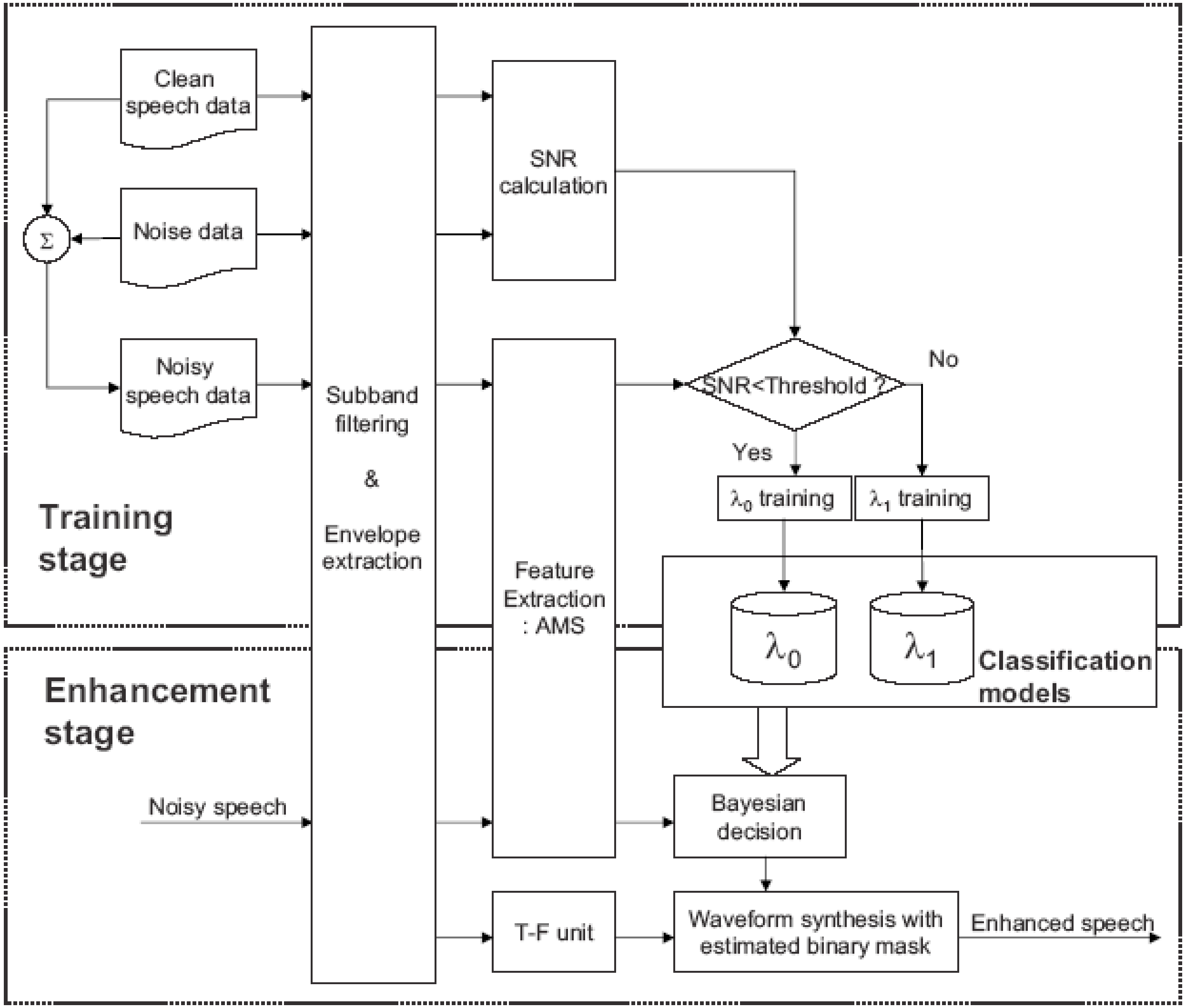
Speech shaped noise



GMM-based classification

- **A classification model by Kim et al. (2009) deals with speech separation in a speaker and masker dependent way:**
  - AM spectrum (AMS) features are used
  - Classification is based on Gaussian mixture models (GMM)
  - Speech intelligibility evaluation is performed with normal-hearing (NH) listeners

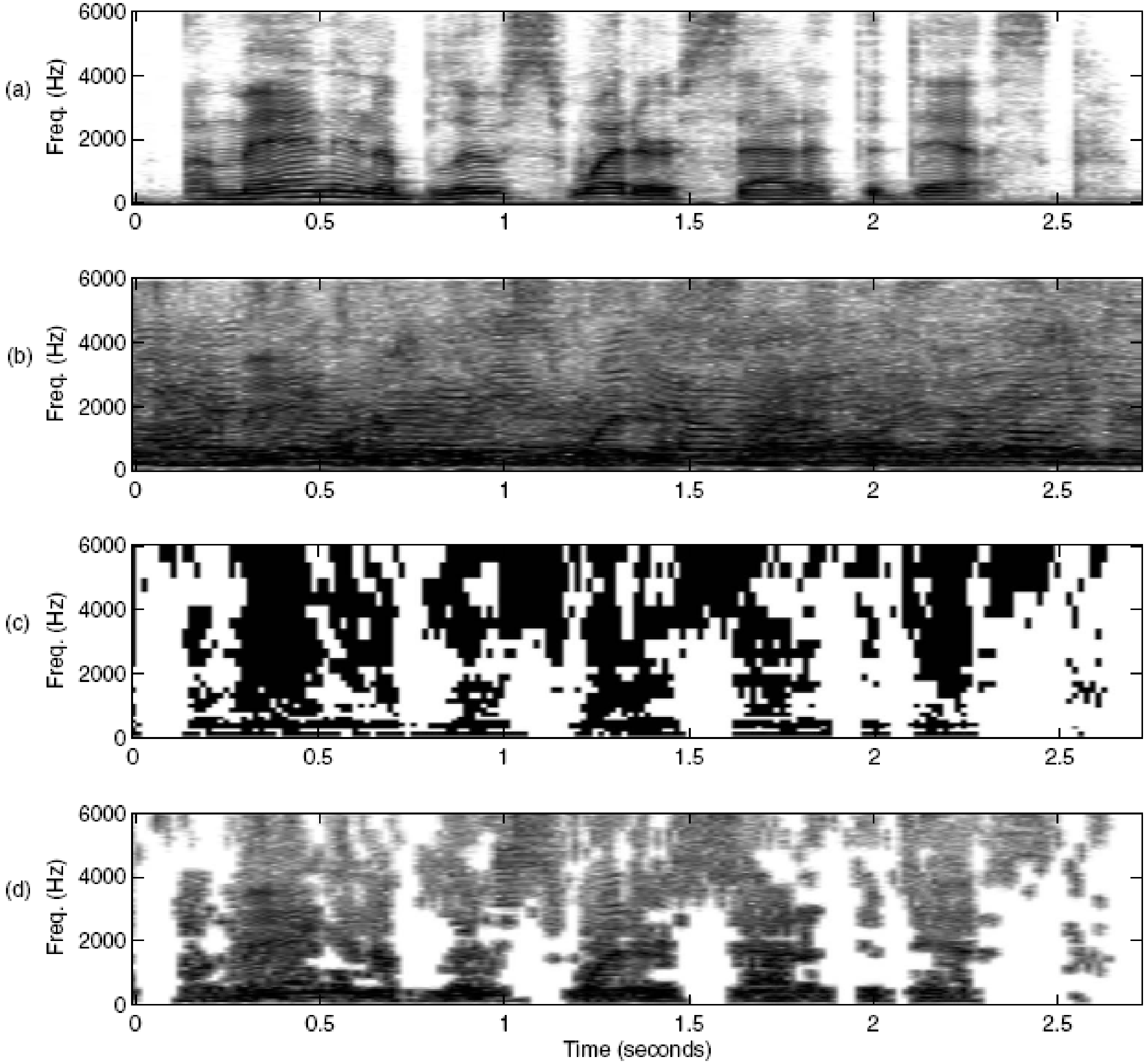
Diagram of Kim et al.’s model



Feature extraction and GMM

- **Peripheral analysis is done by a 25-channel mel-frequency filter bank**
- **An AMS feature vector is extracted within each time-frequency (T-F) unit**
- **The training and test data are mixtures of IEEE sentences and 3 masking noises: babble, factory, and speech-shaped noise**
  - Separate GMMs are trained for each speaker (a male and a female) and each masker

A separation example



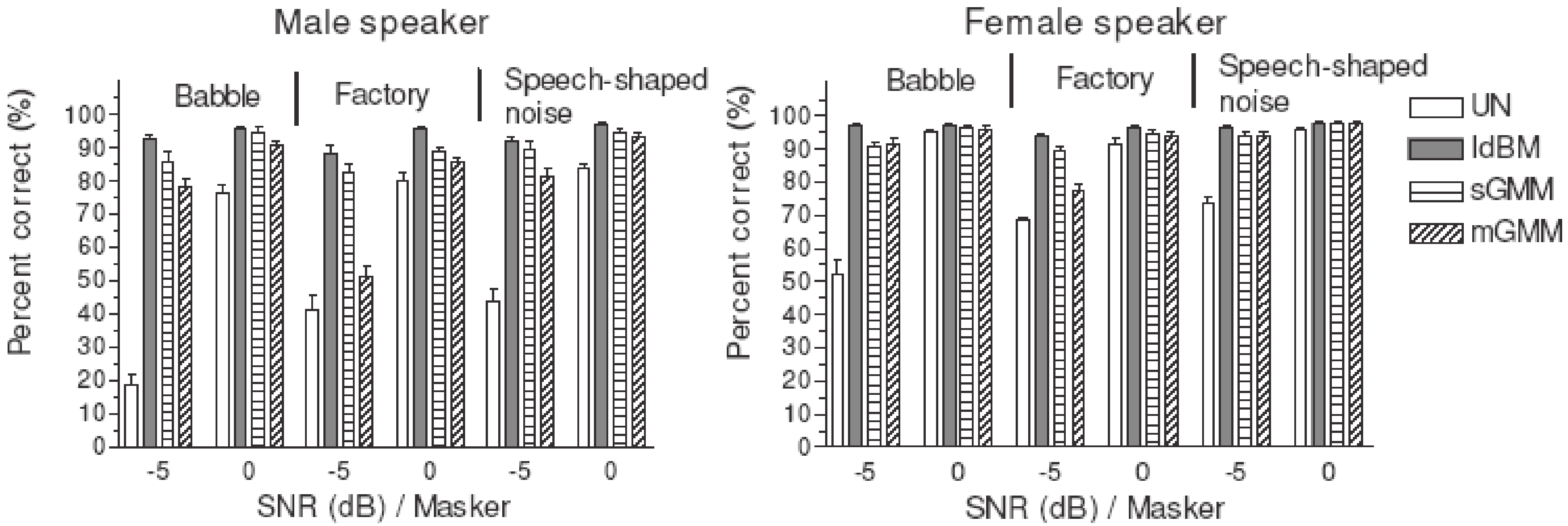
Target utterance

-5 dB mixture with babble

Estimated mask

Masked mixture

Intelligibility results and demo



- First monaural speech segregation algorithm to achieve speech intelligibility improvement

UN: unprocessed
IdBM: ideal binary mask
sGMM: trained on a single noise
mGMM: trained on multiple noises

Clean: 0-dB mixture with babble: Segregated:

Outline of presentation

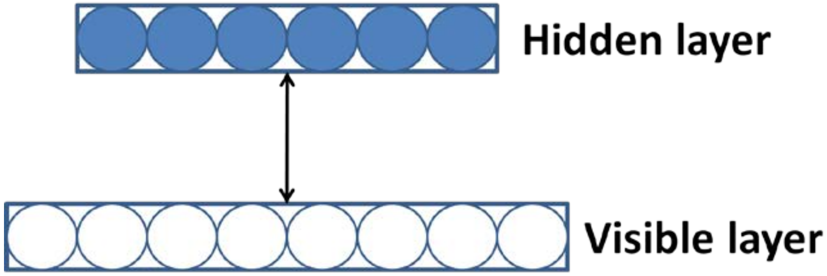
- **Auditory masking and speech intelligibility**
  - Ideal binary mask
  - Separation as binary classification
- **GMM based classification**
  - Speech intelligibility tests on normal hearing listeners
- **DNN based classification**
  - Speech intelligibility tests on hearing impaired listeners
- **Discussion: Problems of SNR**

Deep neural networks (DNNs)

- **Why deep?**
  - Automatically learn more abstract features as the number of layers increases
  - More abstract features tend to be more invariant to superficial variations
  - Superior performance in practice if properly trained (e.g., convolutional neural networks)
- **However, deep structure is hard to train from random initializations**
  - Vanishing gradients: Error derivatives tend to become very small in lower layers, causing overfitting in upper layers
- **Hinton et al. (2006) suggest to unsupervisedly pretrain the network using restricted Boltzmann machines (RBMs)**

Restricted Boltzmann machine

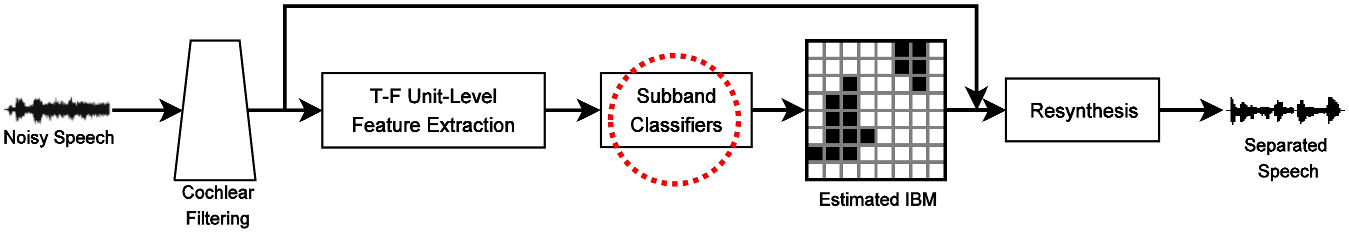
- **RBMs are two-layer (one visible and one hidden layer) networks that model the input distribution**
  - A generative model
- **RBMs simplify Boltzmann machines by allowing connections only between the visible and hidden layer, i.e. no intra-layer recurrent connections**
  - Enables exact and efficient inference



DNN training

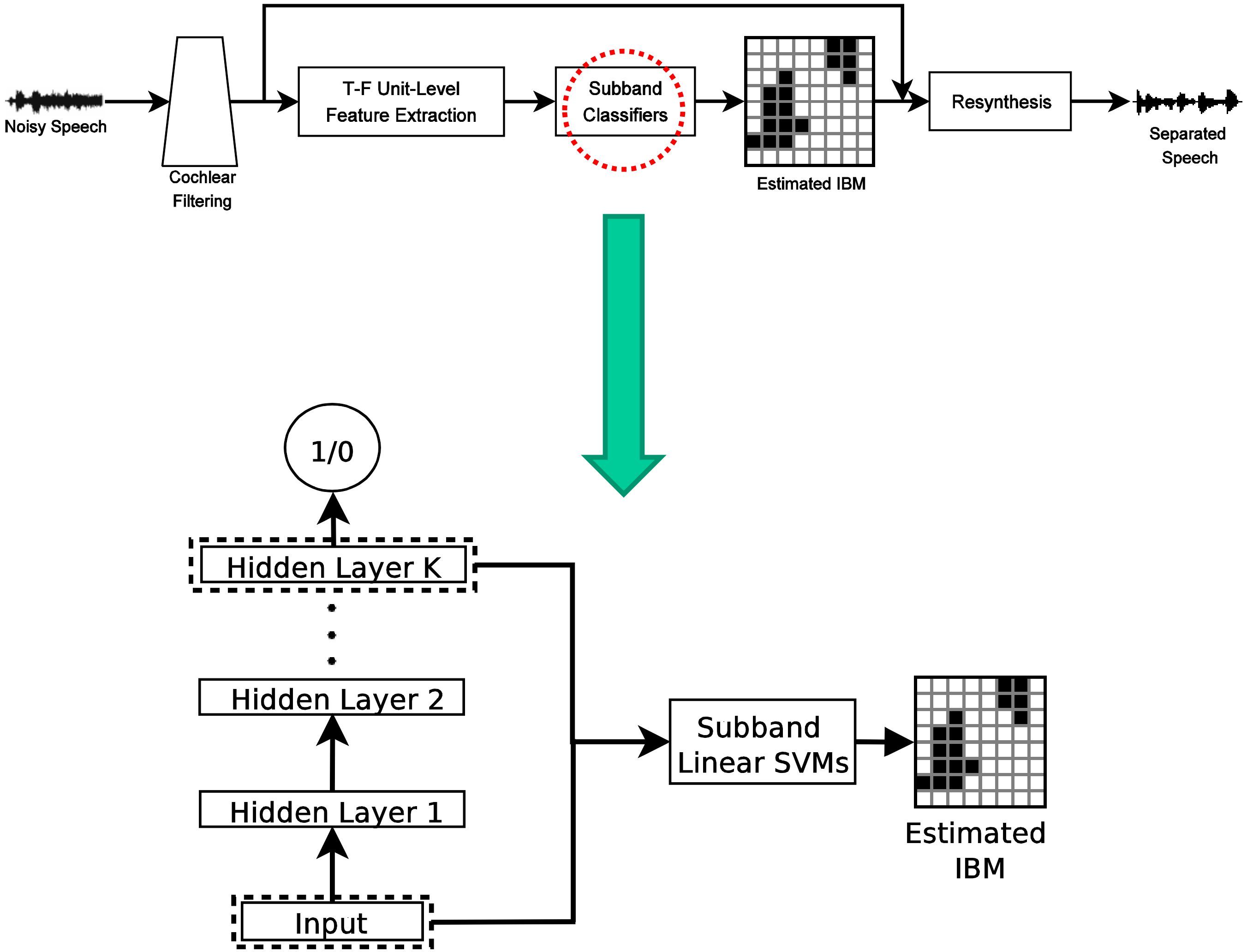
- **Unsupervised, layerwise pre-training via restricted Boltzmann machines**
  - Train the first RBM using unlabeled data
  - Fix the first layer weights. Use the resulting hidden activations as new features to train the second RBM
  - Continue until all layers are thus trained
- **Supervised fine-tuning**
  - The weights from RBM pretraining provide the network initialization
  - Use standard backpropagation (or other discriminative training methods) to fine tune all the weights

DNN as subband classifier (Wang & Wang'13)



- **DNN is used for feature learning from raw acoustic features**
  - Train DNNs in the standard way. After training, take the last hidden layer activations as learned features
- **Train SVMs using the combination of raw and learned features**
- **Linear SVM seems adequate**
  - The weights from the last hidden layer to the output layer essentially define a linear classifier
  - Therefore the learned features are amenable to linear classification

DNN as subband classifier (Wang & Wang'13)



Noisy Speech

Cochlear Filtering

T-F Unit-Level Feature Extraction

Subband Classifiers

Estimated IBM

Resynthesis

Separated Speech

1/0

Hidden Layer K

Hidden Layer 2

Hidden Layer 1

Input

Subband Linear SVMs

Estimated IBM

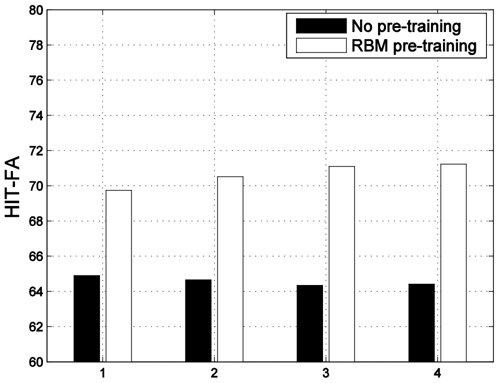
DNN pilot study

- **Since this is the first DNN study for separation, we first train on a small corpus: 50 sentences mixed with 12 noises at 0 dB**
- **Raw feature: RASTA-PLP + Delta + Acceleration (39-D)**

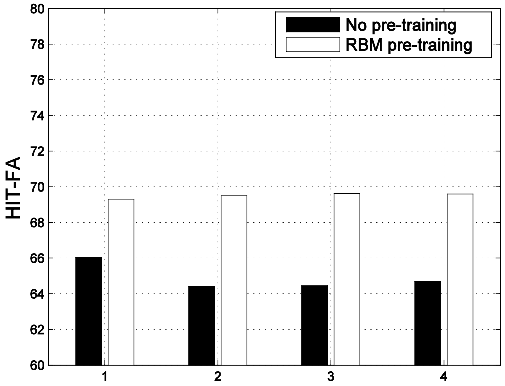
System	Overall HIT-FA	Voiced HIT-FA	Unvoiced HIT-FA
Linear SVM	56.5%	63.0%	34.5%
Gaussian SVM	68.7%	73.4%	51.5%
DNN-SVM	73.2%	75.3%	64.6%
DNN-gSVM	74.3%	76.5%	66.0%

- **Linear SVMs on learned features are much better than on raw features**
- **DNN-SVM outperforms kernel SVM significantly, especially in unvoiced intervals**
    - With learned features, kernel SVM (gSVM) with high complexity only produces marginal improvements

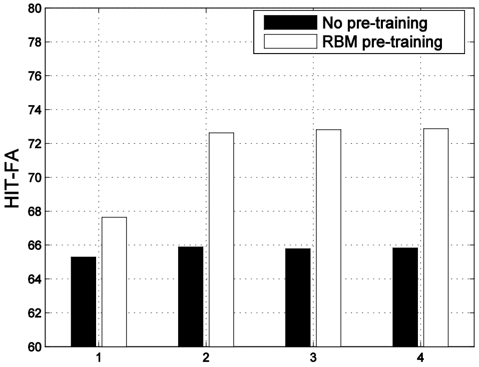
Effects of RBM Pretraining



Matched-noise condition



Unmatched-noise condition



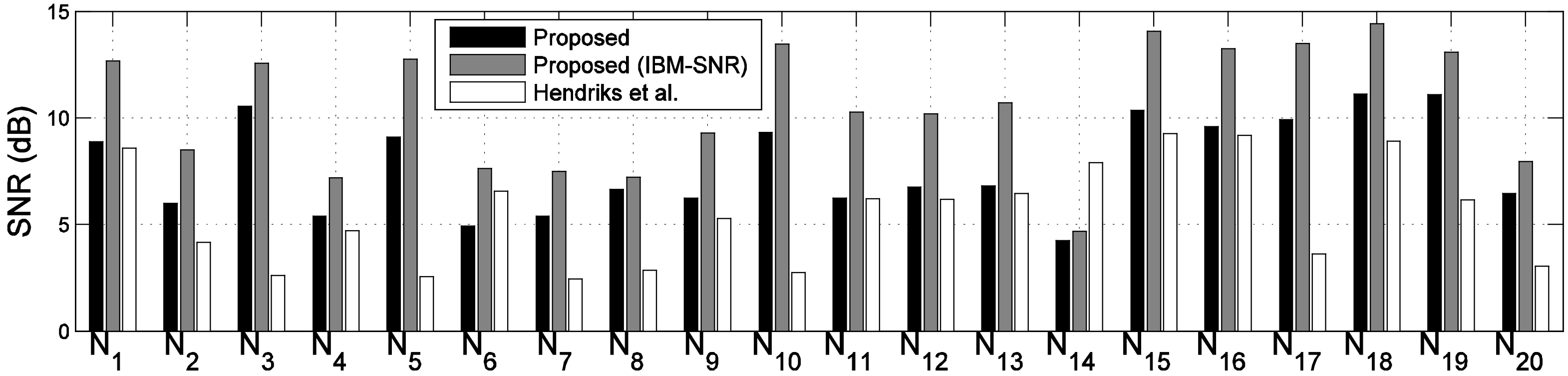
-5 dB SSN/Babble noise (more challenging)

HIT-FA comparisons as a function of the number of hidden layers

Extensive training with DNN

- **Training on 200 randomly chosen utterances from both male and female IEEE speakers, mixed with 100 environmental noises (Hu’04) 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 separation results



- Comparisons with a representative speech enhancement algorithm (Hendriks et al. 2010)
- Using clean speech as ground truth, on average about 3 dB SNR improvements
- Using IBM separated speech as ground truth, on average about 5 dB SNR improvements

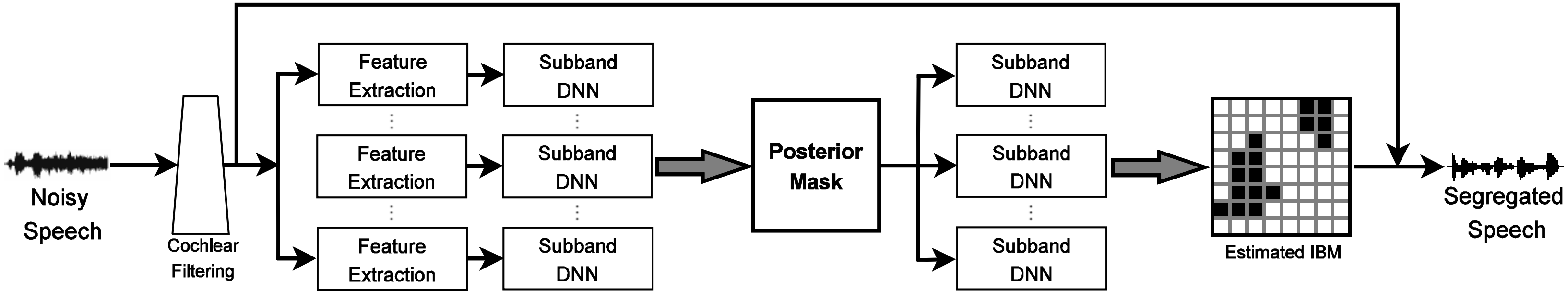
Sound demos

Speech mixed with unseen, daily noises

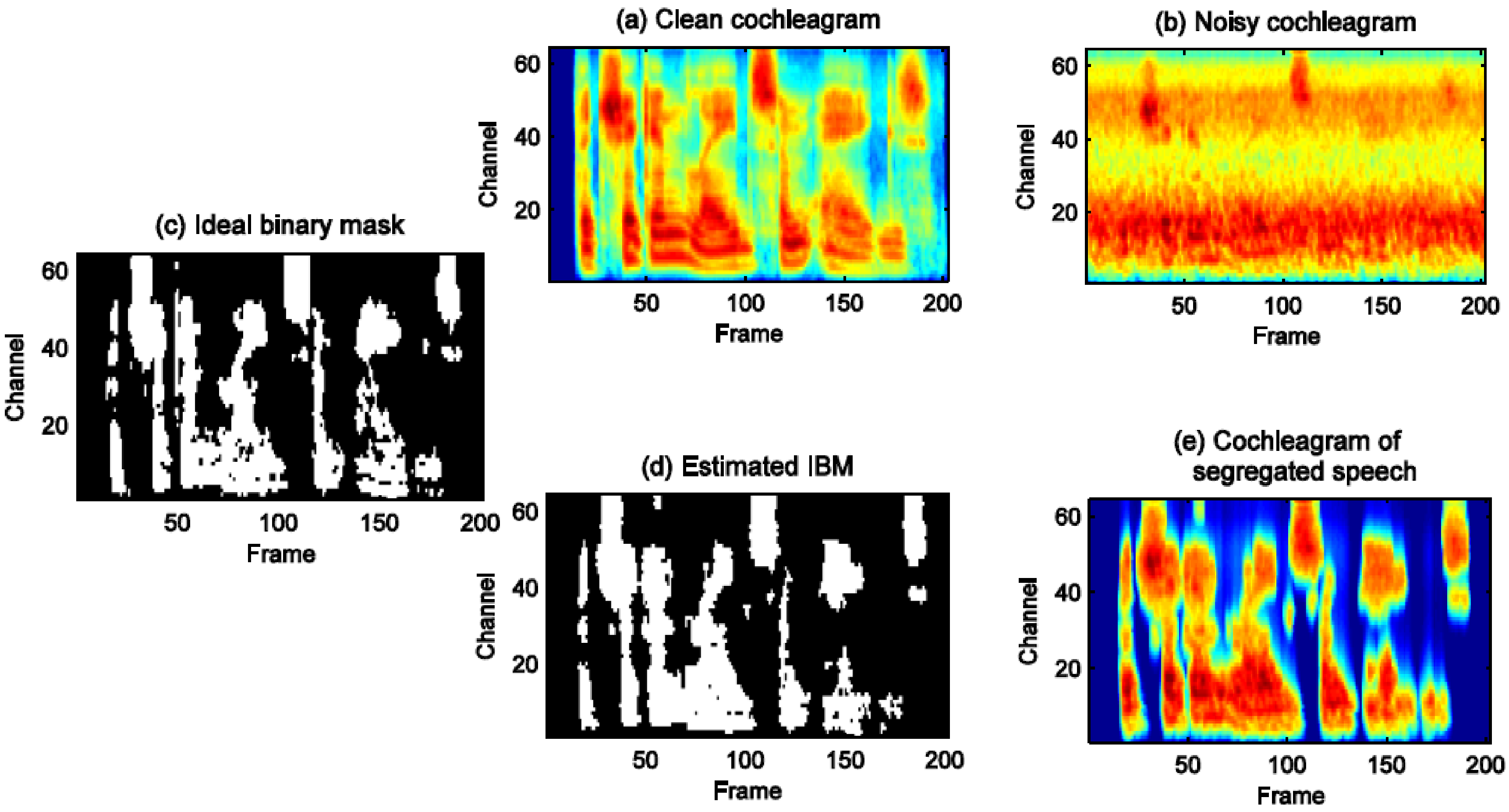
Cocktail party noise (5 dB)	🔊 Mixture	🔊 Separated
Destroyer noise (0 dB)	🔊 Mixture	🔊 Separated

Speech intelligibility evaluation

- **We recently tested speech intelligibility of hearing-impaired (HI) listeners (Healy et al.'13)**
  - A very challenging problem: "The interfering effect of background noise is the single greatest problem reported by hearing aid wearers" (Dillion'12)
- **Two stage DNN training to incorporate T-F context in classification**

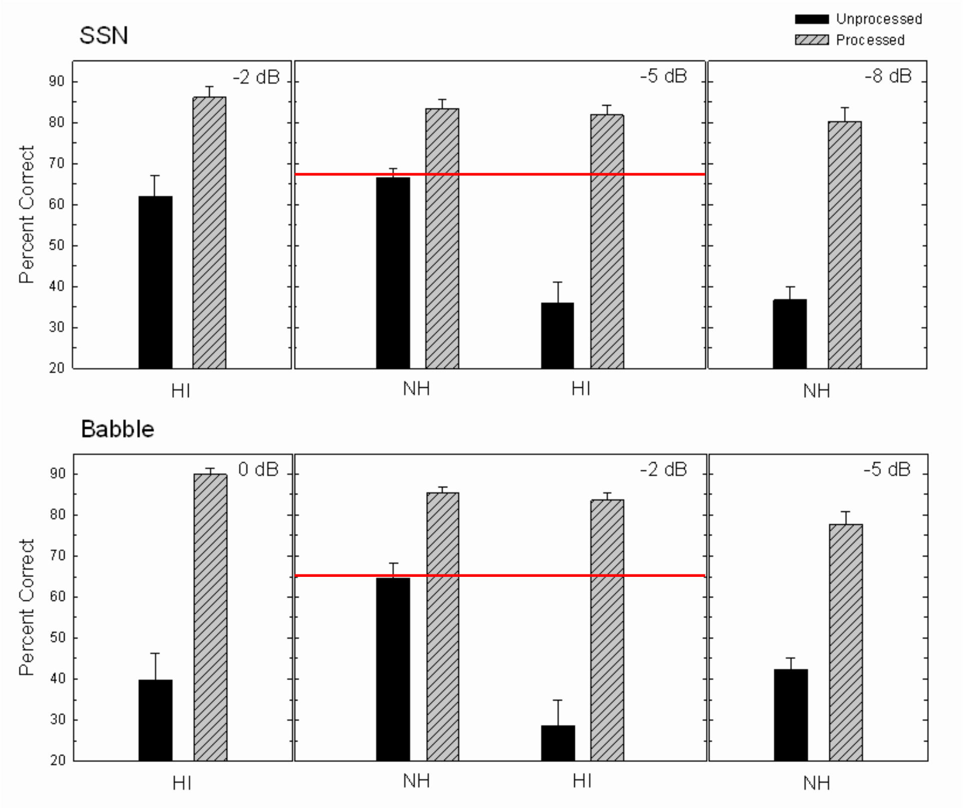


An illustration



A HINT sentence mixed with speech-shaped noise at -5 dB SNR

Results and sound demos



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

Discussion: Problems of SNR

- **The SNR is probably the most commonly used performance metric for speech separation/enhancement**
- **SNR maximization aims to produce an output signal as close to the target signal as possible**
- **In binary masking, however, negative LC values are needed for higher intelligibility (Brungart et al.’06, Li & Loizou’08, Kim et al.’09, Healy et al.’13)**
  - This is to retain some speech underneath noise
- **Compared to 0 dB, a negative LC leads to lower SNR**
  - That is, lower SNR yields higher intelligibility

Discussion: Problems of SNR (cont.)

- **The SNR metric does not distinguish amplification distortion and attenuation distortion**
  - But, amplification distortion is much more detrimental to intelligibility (Loizou & Kim'11)
  - Similarly, false alarm error in binary masking is much more detrimental than miss error (Li & Loizou'08)
- **Widespread use of SNR (or its variants) to evaluate enhancement/separation is partly responsible for lack of intelligibility improvement**
  - What's the alternative? HIT-FA (Kim et al.'09), STOI (Taal et al.'11)?

Conclusion

- **From auditory masking to the IBM notion, to binary classification for speech separation**
  - In other words, separation is about classification, not target estimation
- **This new formulation enables the use of supervised learning**
  - Extensive training with DNN is a promising direction
- **This approach has yielded the first demonstrations of speech intelligibility improvement in noise**

Acknowledgments

- **Funding provided by NIDCD and AFOSR**