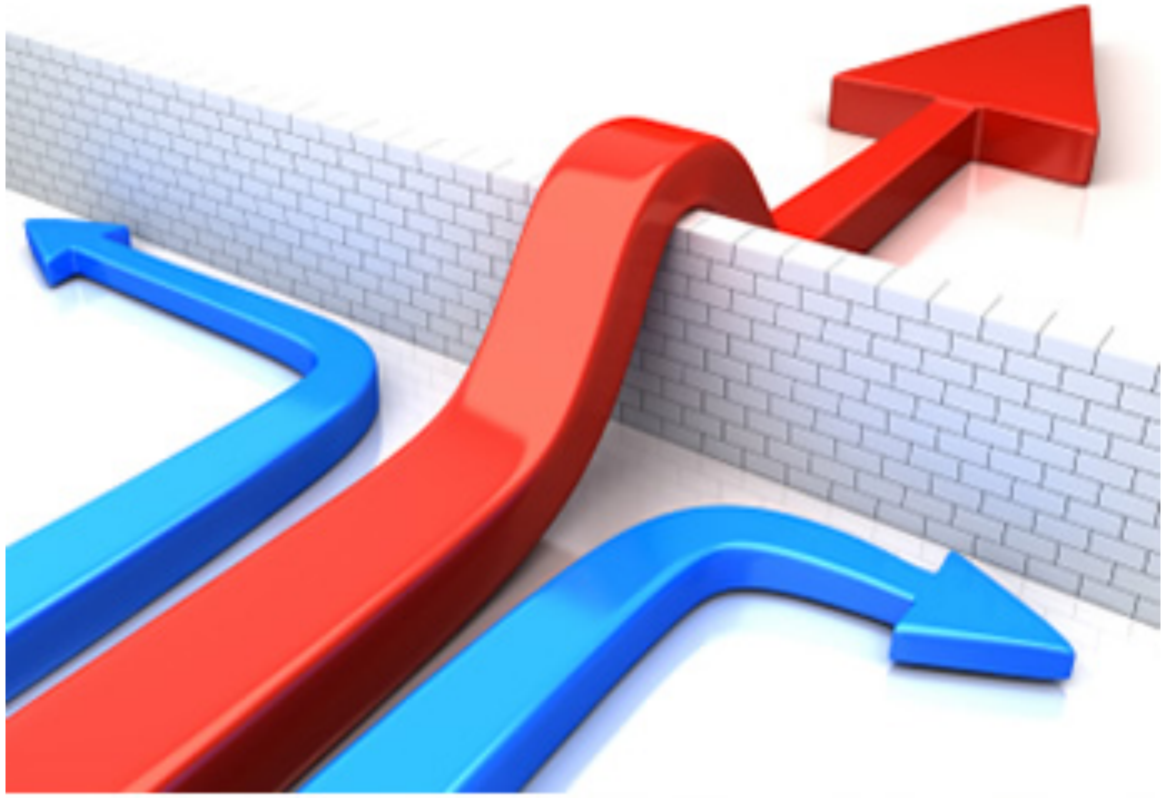
An algorithm to improve speech recognition in noise for hearing impaired listeners

E. W. Healy, S. E. Yoho, Y. Wang,
D. Wang
Presented by Sara Sabour

Challenges

- Noisy background
- Monaural input
- No prior knowledge
- Hearing impaired listener

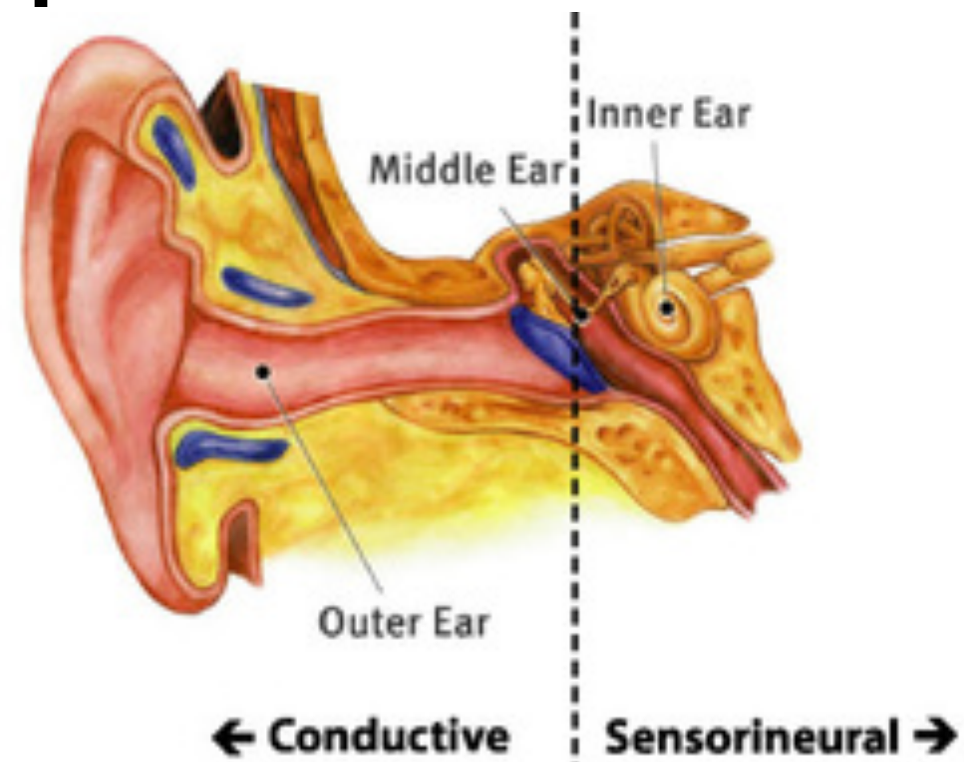


Overview

- Related works
- Algorithm
  - Feature extraction
  - Training
- Testing
- Results

Hearing Impaired

- Reduced audibility
- Reduces frequency resolutior
- Across frequency deficiency



Related Works

- Microphone arrays, limits
  - Assumption of different spatial locations
  - Configuration stationarity
- Single Microphone, statistical analysis limits
  - Lack of increase in intelligibility for human due to
    - Musical noise
    - Removal of low intensity sounds

Related Work

- IBM: ideal binary time frequency mask
  - Matrix with 1 for each TF in which SNR>T
  - Ideal: Prior knowledge and optimal SNR gain
- Without prior knowledge
  - How to estimate? Kim et al. (2009)
    - Gaussian mixture model
    - For normal hearing
    - Adopted for cochlear implant users
    - GMM overfit

Overview of Algorithm

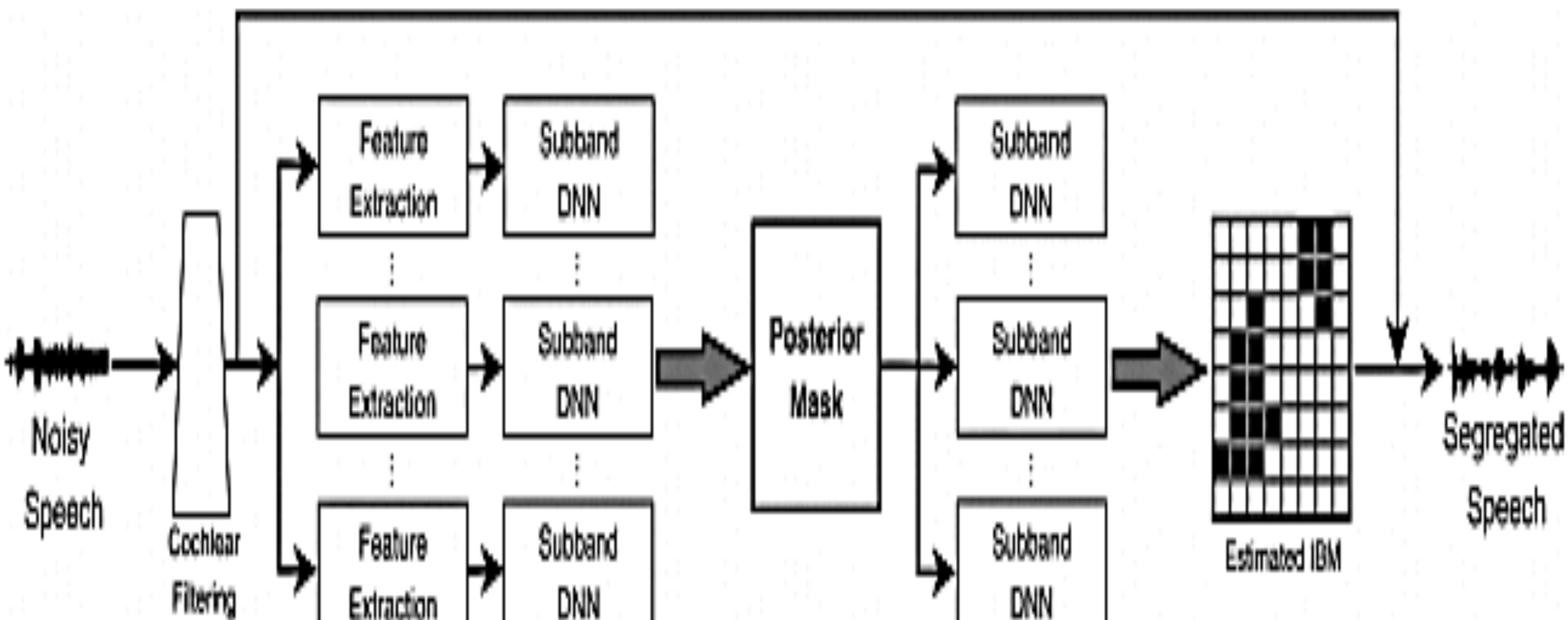
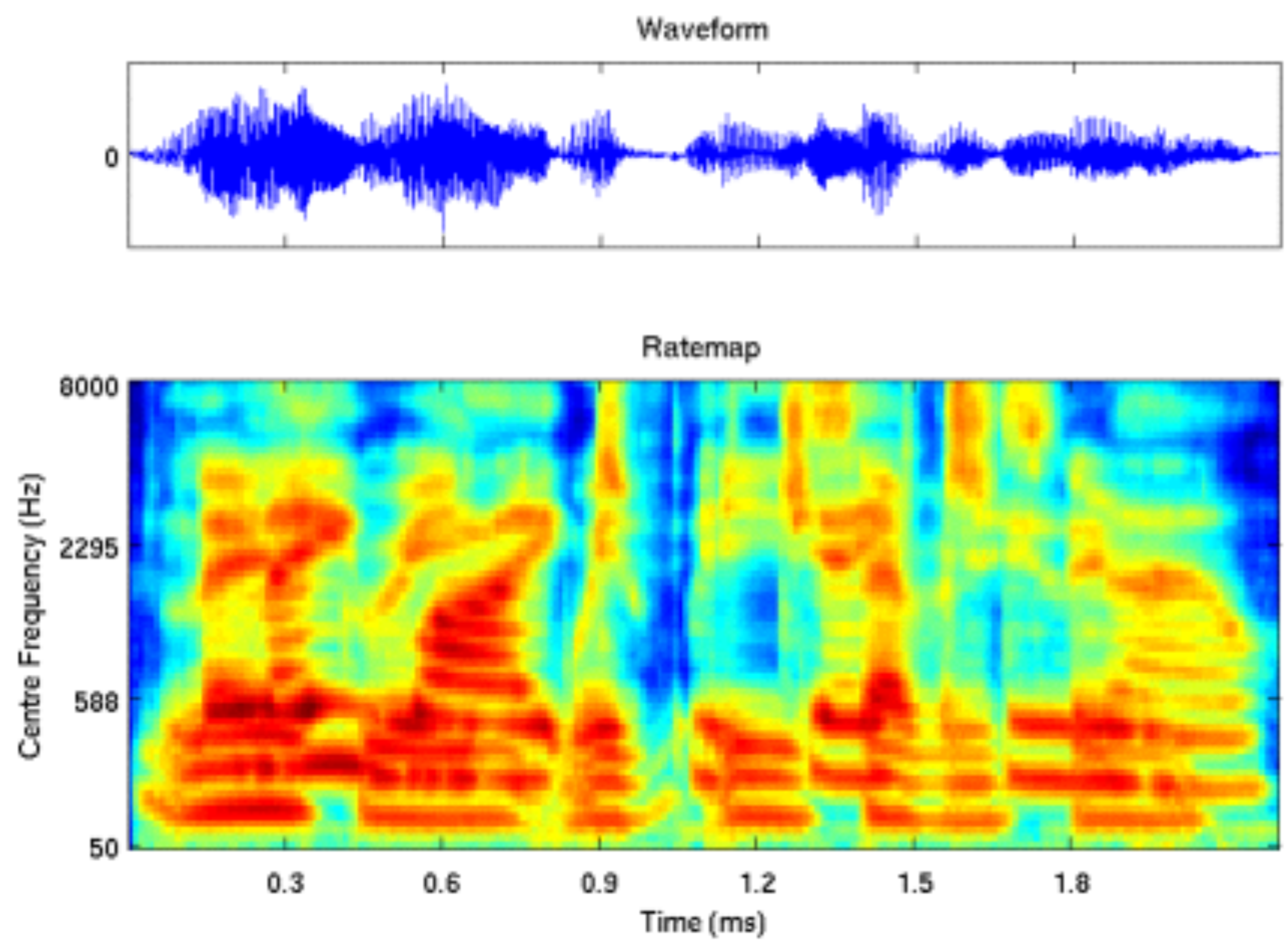


FIG. 1. Schematic diagram of the current speech-segregation system. DNN = deep neural network, IBM = ideal binary mask.

Filtering

- 64 channel, 50 to 8000 Hz, 20ms with 10ms overlap
- Gammatone
- Cochleagram



Feature Extraction

- Amplitude modulation spectrogram(AMS)
- Relative spectral transform and preceptual linear prediction(RASTA-PLP)
- Mel-frequency cepstral coefficients(MFCC)
- Delta features on RASTA-PLP
- 85-D feature vector

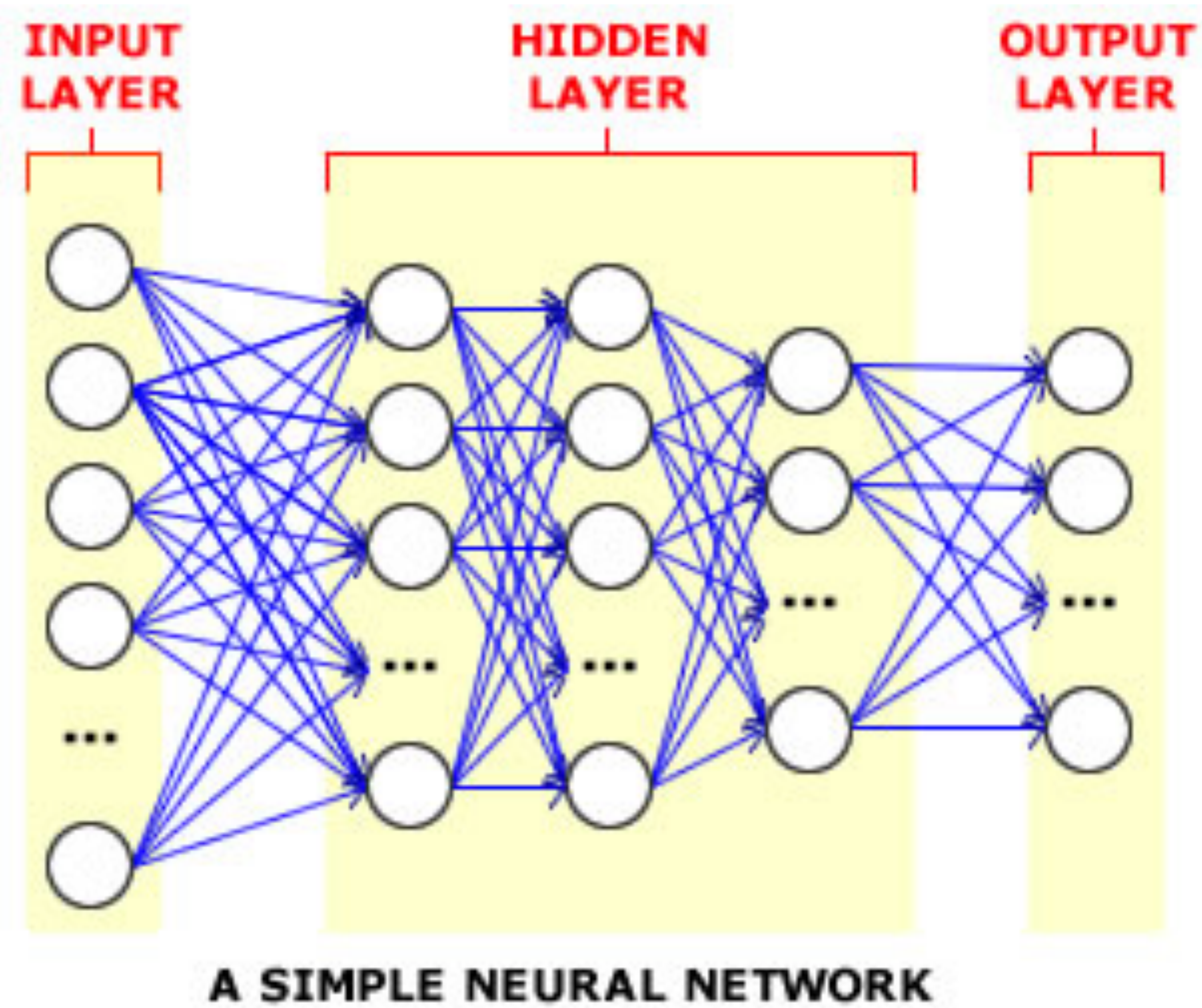
Classifier

- Deep neural networks within each frequency
- Two hidden layers & RBM pretraining
- Gaussian-Bernoulli and Bernoulli-Bernoulli RBMs
- 200 units and sigmoid transfer function
- Cross entropy objective function error of IBM
- Mini-batch gradient descent, batch size 512

$$L(\mathbf{w}) \ = \ -\frac{1}{N}\sum_{n=1}^{N} H(p_n, q_n) \ = \ -\frac{1}{N}\sum_{n=1}^{N} \ \left[y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n)\right]$$

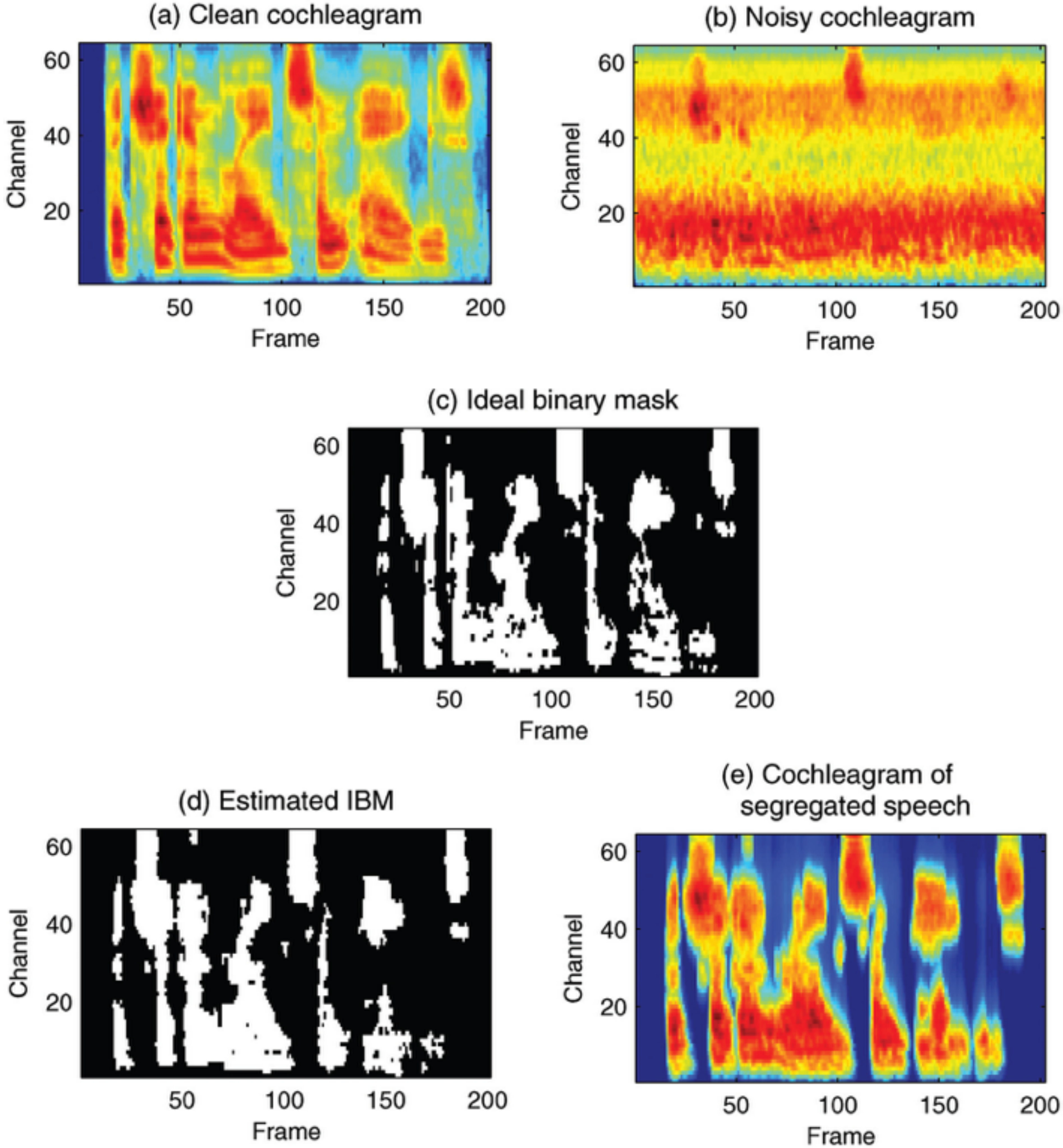
Deep Neural Networks

1. Initialize all the weights
2. Send the input through
3. Calculate the loss
4. Back propagate
5. Update the weights
6. Iterate



Contextual Information

- Capturing structured spectro-temporal patterns
- Idea: concatenating neighbouring TFs posterior probability
- Window of 5 time frame and 17 frequency channel


(a) Clean cochleagram
(b) Noisy cochleagram
(c) Ideal binary mask
(d) Estimated IBM
(e) Cochleagram of segregated speech
Channel
Frame
60
40
20
50
100
150
200

Overview of Algorithm

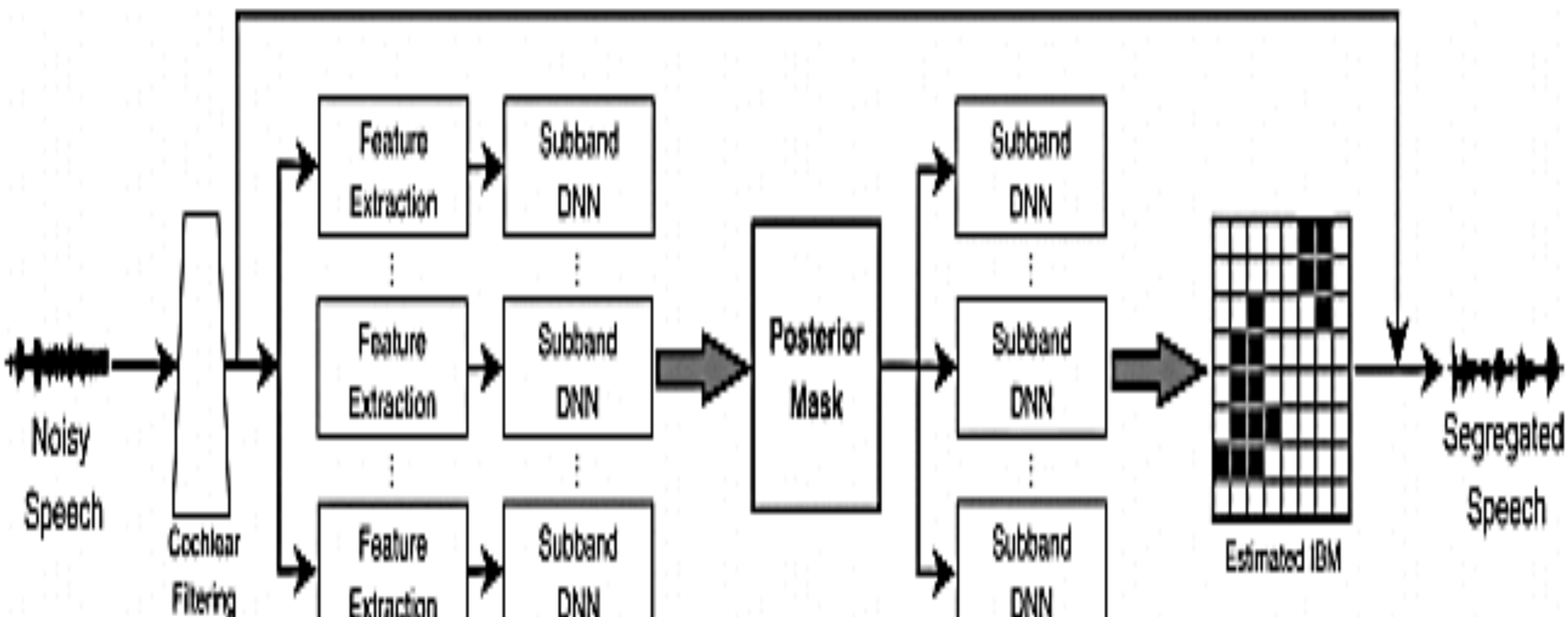
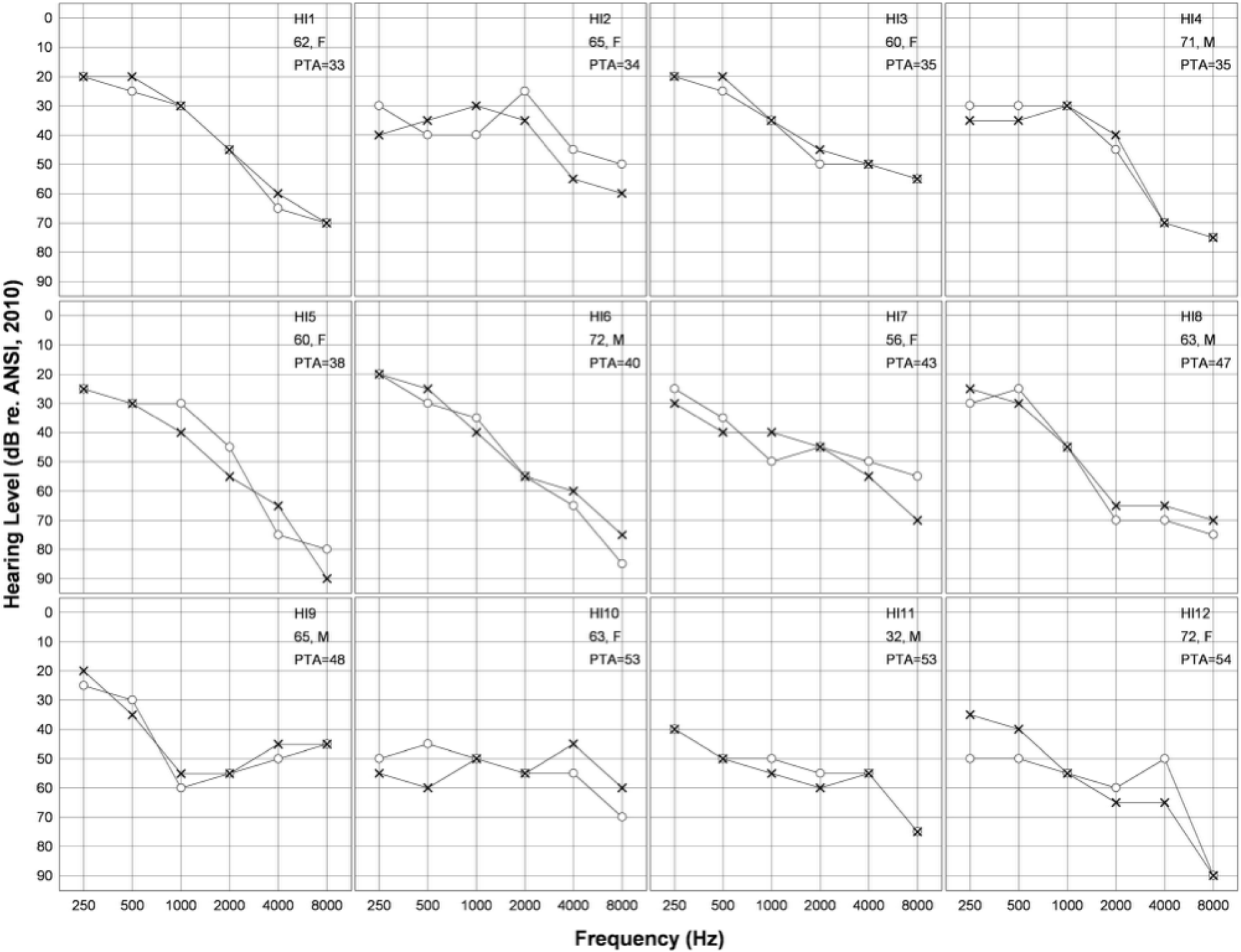


FIG. 1. Schematic diagram of the current speech-segregation system. DNN = deep neural network, IBM = ideal binary mask.

Testing: Subjects

- 12 NH listeners
  - aged 19-28 (mean = 21)
  - 20 dB at octave frequencies from 250 to 8000 Hz
  - Female
- 12 diagnosed with bilateral sensorineural hearing loss of cochlear origin

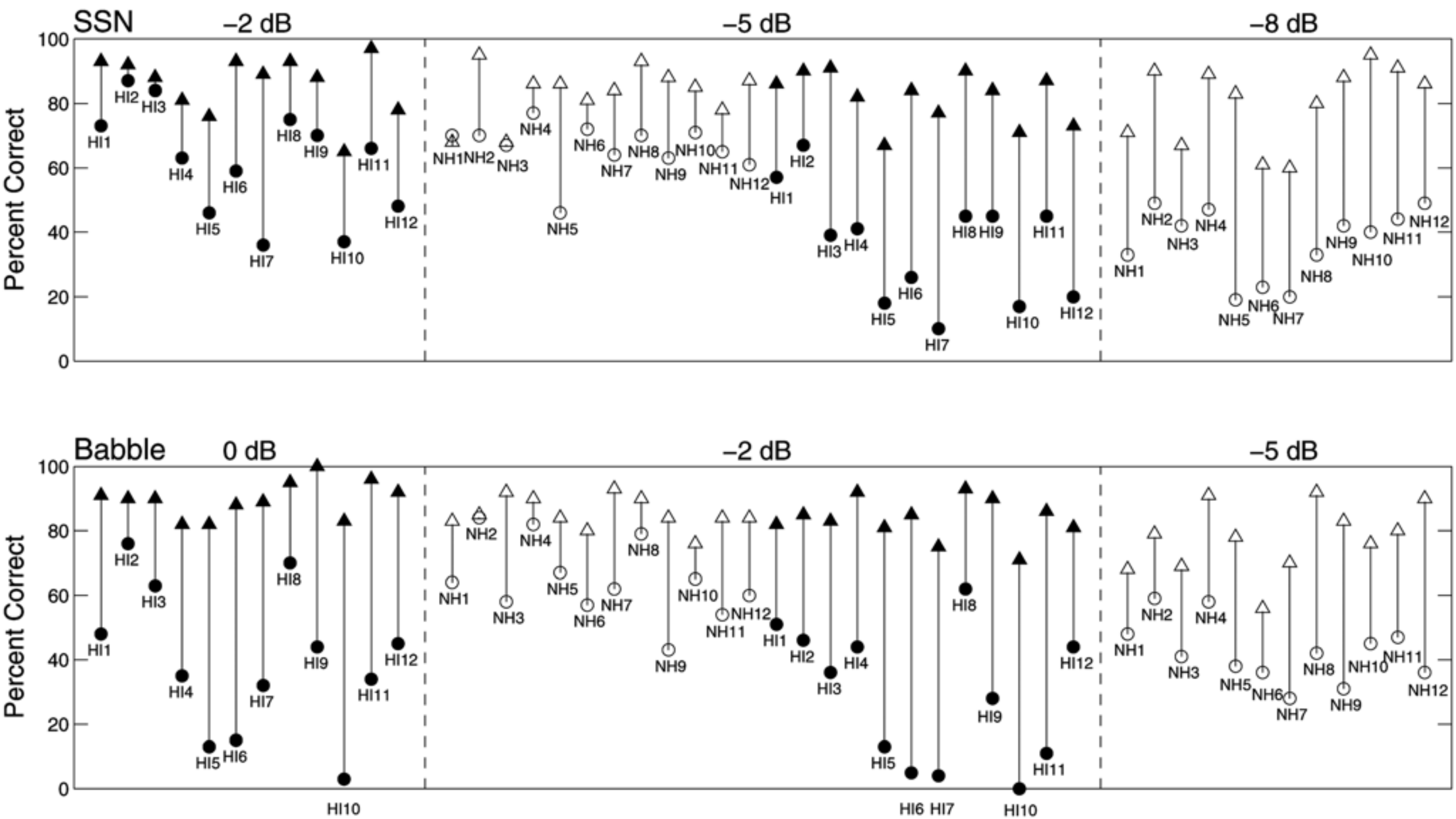

HI1
62, F
PTA=33
HI2
65, F
PTA=34
HI3
60, F
PTA=35
HI4
71, M
PTA=35
HI5
60, F
PTA=38
HI6
72, M
PTA=40
HI7
56, F
PTA=43
HI8
63, M
PTA=47
HI9
65, M
PTA=48
HI10
63, F
PTA=53
HI11
32, M
PTA=53
HI12
72, F
PTA=54
Hearing Level (dB re. ANSI, 2010)
0
10
20
30
40
50
60
70
80
90
250
500
1000
2000
4000
8000
Frequency (Hz)

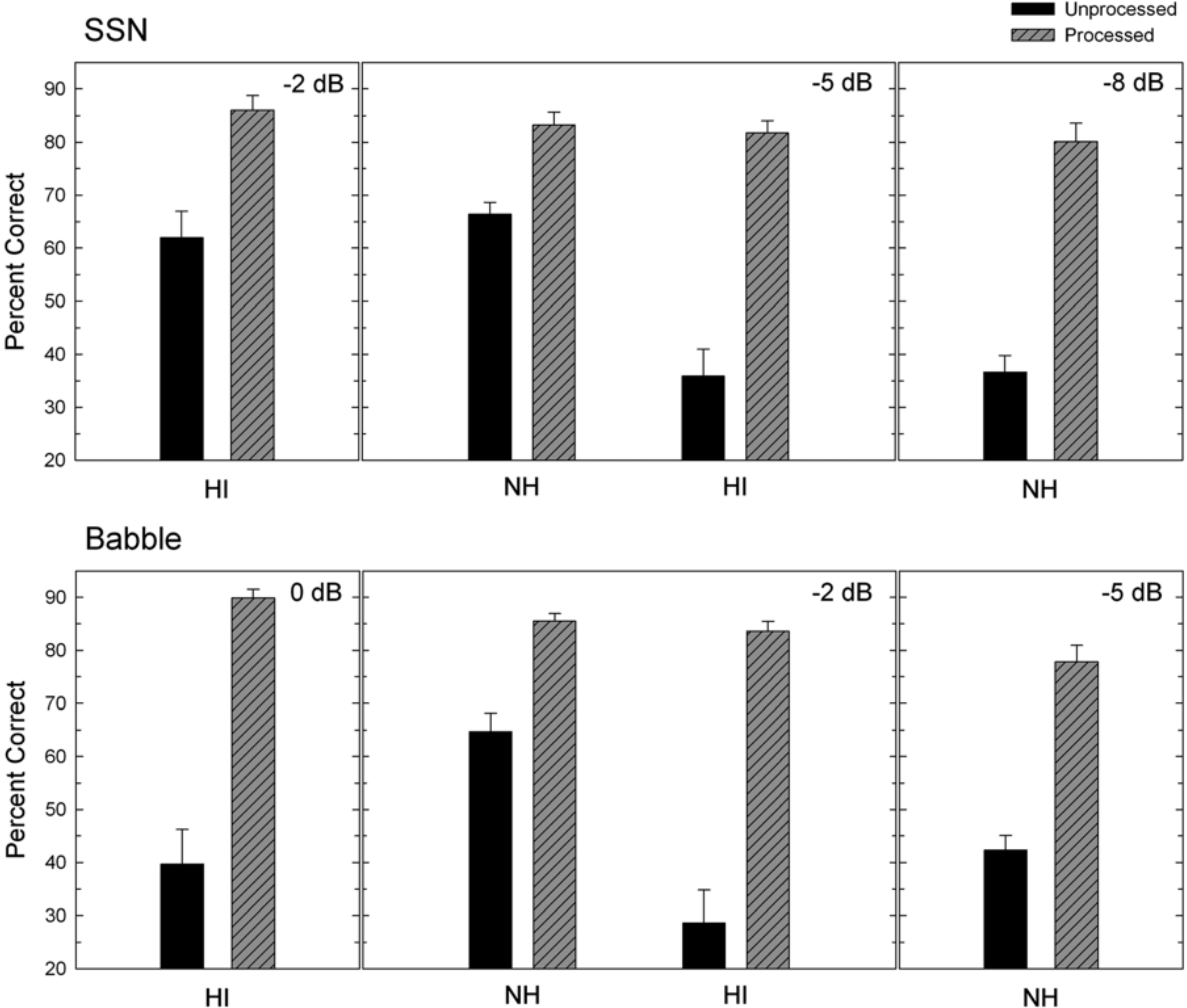
Data

- Male talker recordings of HINT sentences (Nilsson et al.)
- 16kHz and **equal RMS** energy
- Different training and test data
- Random start point looped **Noise**, 140 ms margin of original signal
    - **SSN**: commercial HINT, 10s
        - -2, -5, -8 dB SNR (over 1 HINT)
    - **babble**: 8 talkers of TIMIT (Garofolo et al, 1993)
        - 0, -2, -5 dB SNR (over 1 HINT)

Procedure

- Familiarization
  - 5 sentences in each: quiet UP, SSN UP, SSN P, babble UP, babble P
  - 0 dB SNR
- Testing
  - 20 HINT, 8 conditions
    - (2 P/UP X 2 SSN/Babble X 2 SNRs)
      - Each HI/NH: 2/3 SNRs
    - Pseudo-randomized condition order, list order
      - UP/P successively and random
  - RMS
    - NH: 65 dBA
    - HI: 85 dBA, 1: 90 dBA


SSN
−2 dB
−5 dB
−8 dB
Percent Correct
Babble
0 dB
−2 dB
−5 dB
Percent Correct
HI1
HI2
HI3
HI4
HI5
HI6
HI7
HI8
HI9
HI10
HI11
HI12
NH1
NH2
NH3
NH4
NH5
NH6
NH7
NH8
NH9
NH10
NH11
NH12



Unprocessed
Processed
SSN
Percent Correct
-2 dB
-5 dB
-8 dB
HI
NH
HI
NH
Babble
0 dB
-2 dB
-5 dB
HI
NH
HI
NH
20
30
40
50
60
70
80
90

Compare to Kim et al.

- Both: utility of binary classification
- Amount of improvement
    - Different speech material and noise
- Quality of IBM estimation
    - HIT-FA: percent of correctly classified and percent of false alarms
    - SSN -5 dB: 79.3% vs (64.2% & 76.1%)
    - Babble -5 dB: 80.9% vs (59.4% & 72.4%)

Future Work

- Fact: Intelligibility by HI exceeds NH
    - Simplified feature extraction
    - Optimized Matlab code
    - Technology oriented implementation
- Generalization
    - Talker: Not an issue
    - SNR level: Not a Major issue
    - Noise type
        - Han and Wang: training on absent frames
        - Wang and Wang: large number of noise training