Monaural and Binaural Speech Separation

DeLiang Wang

Perception & Neurodynamics Lab

The Ohio State University

Outline of presentation

- **Introduction**
  - CASA approach to sound separation
  - Ideal binary mask as CASA goal
- **Voiced speech separation based on pitch tracking and amplitude modulation analysis**
- **Unvoiced speech separation based on onset/offset analysis**
- **Binaural separation based on sound localization**
- **Summary and discussion**

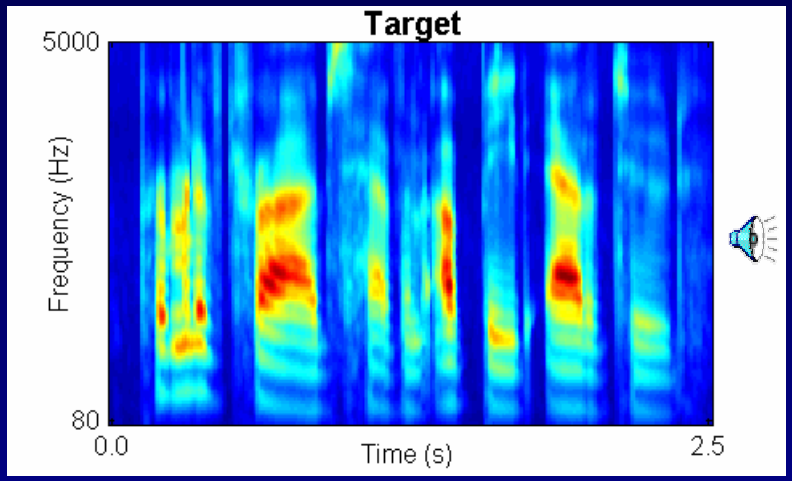
Auditory scene analysis

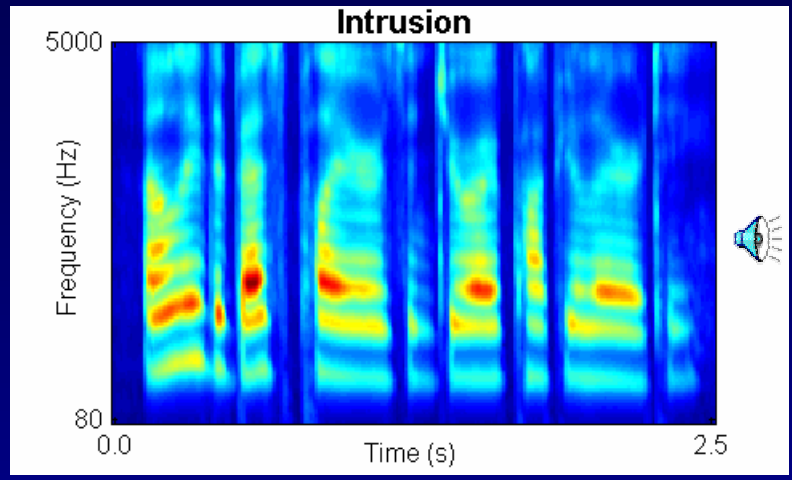
- **The auditory system shows a remarkable capacity in monaural segregation of sound sources in the perceptual process of auditory scene analysis (ASA)**
- **ASA takes place in two conceptual stages (Bregman'90):**
  - **Segmentation**. Decompose the acoustic signal into segments (sensory elements)
  - **Grouping**. Combine segments into streams so that the segments of the same stream likely originate from the same source
- **Computational ASA (CASA) approaches sound separation based on ASA principles**

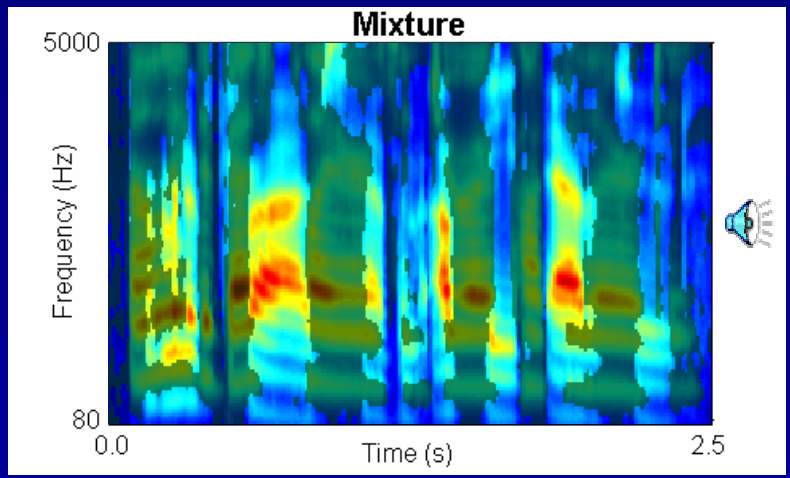
Ideal binary mask as CASA goal

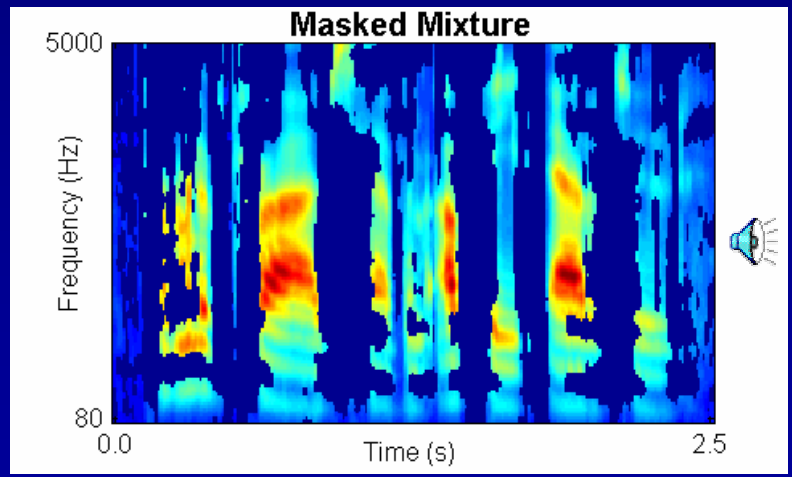
- **Key idea is to retain parts of a target sound that are stronger than the acoustic background, or to mask interference by the target**
  - Broadly consistent with auditory masking and speech intelligibility results
- **Within a local time-frequency (T-F) unit, the ideal binary mask is 1 if target energy is stronger than interference energy, and 0 otherwise**
  - Local 0-dB SNR criterion for mask generation

Ideal binary mask illustration





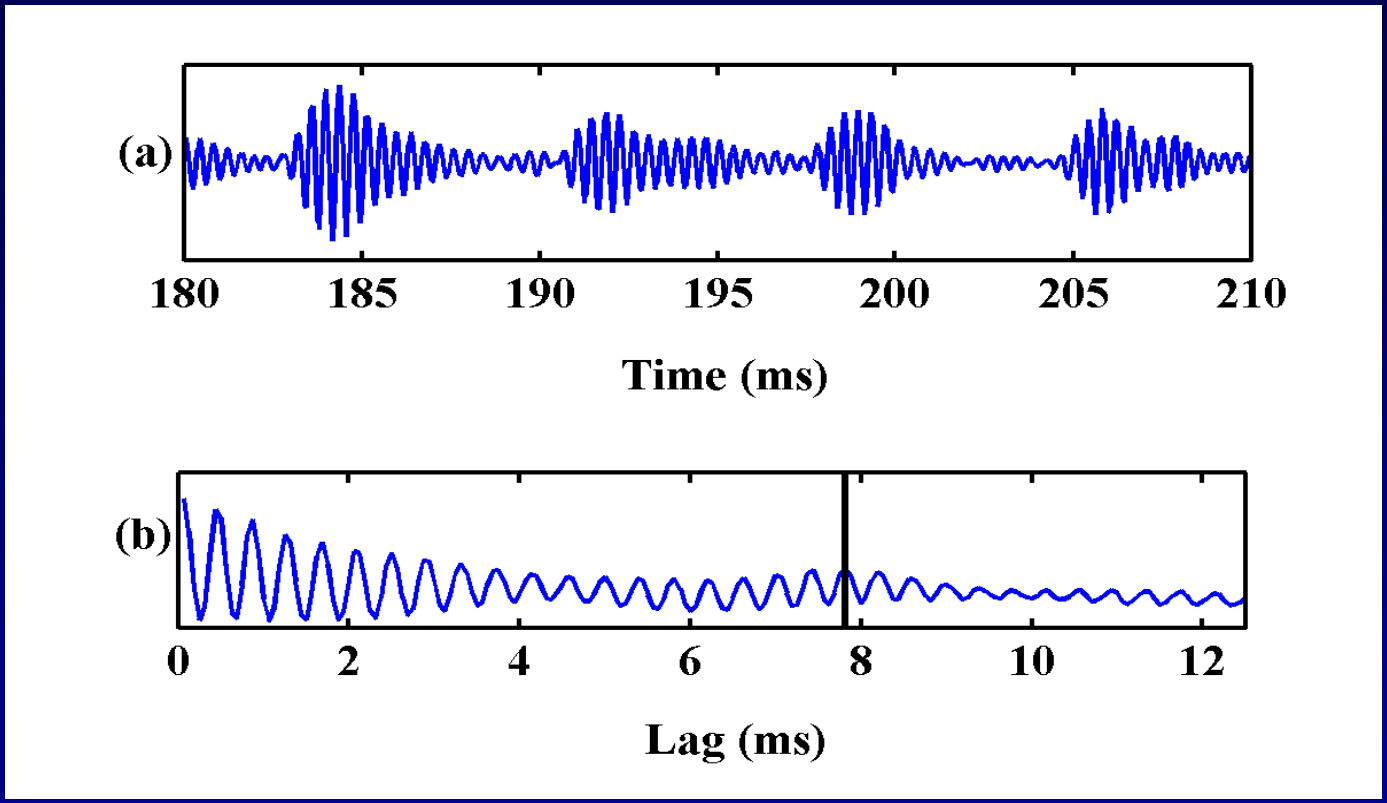




Monaural segregation of voiced speech

- **For voiced speech, lower harmonics are resolved while higher harmonics are not**
- **For unresolved harmonics, a filter channel responds to multiple harmonics, and its response is amplitude modulated (AM)**
- **A CASA model by Hu & Wang (2004) applies different grouping mechanisms in the low-frequency and high-frequency ranges**
  - Low-frequency signals are grouped based on periodicity and temporal continuity
  - High-frequency signals are grouped based on AM and temporal continuity

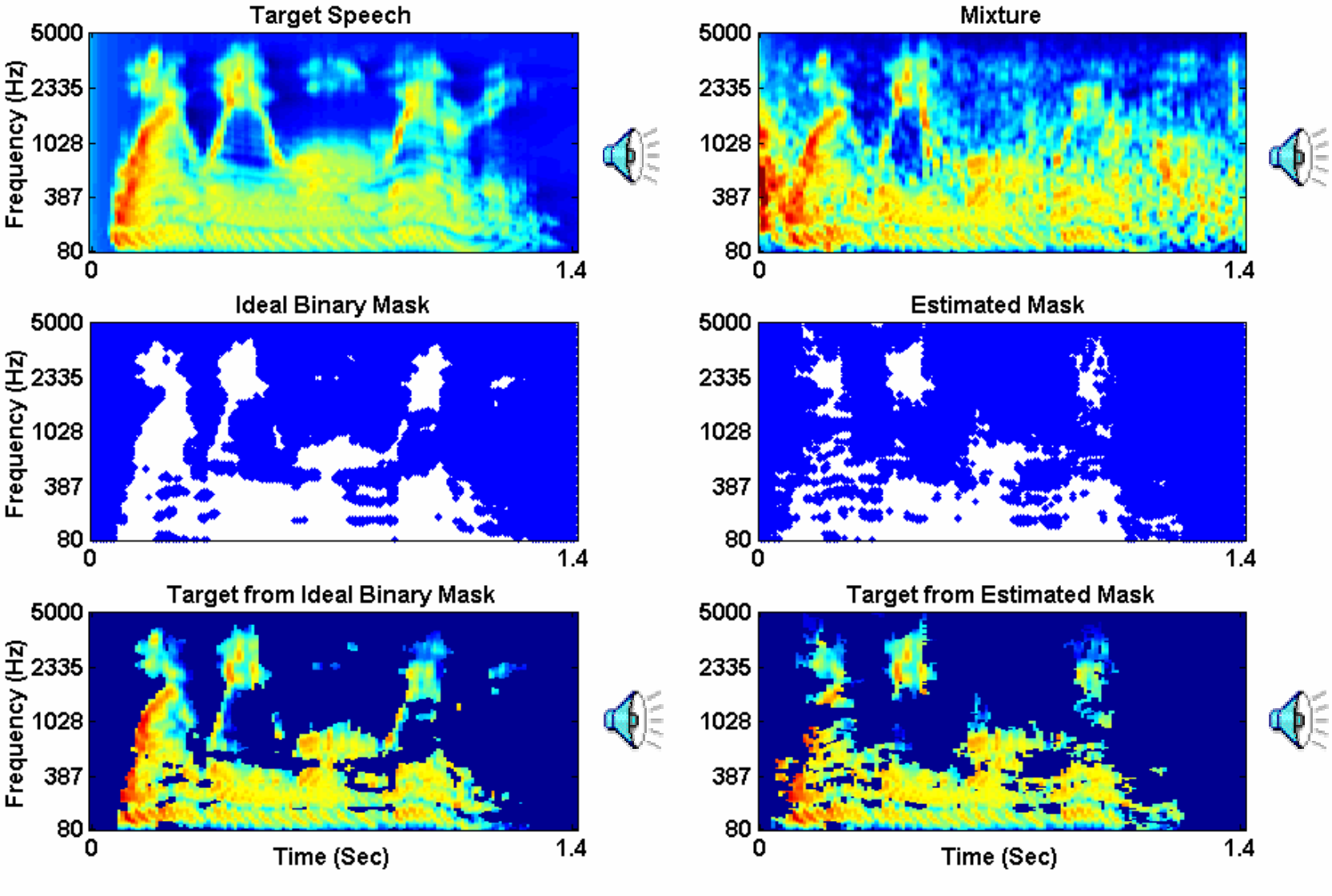
AM illustration



(a) The output of a gammatone filter (center frequency: 2.6 *k*Hz) in response to clean speech

(b) The corresponding autocorrelation function

Voiced speech segregation example



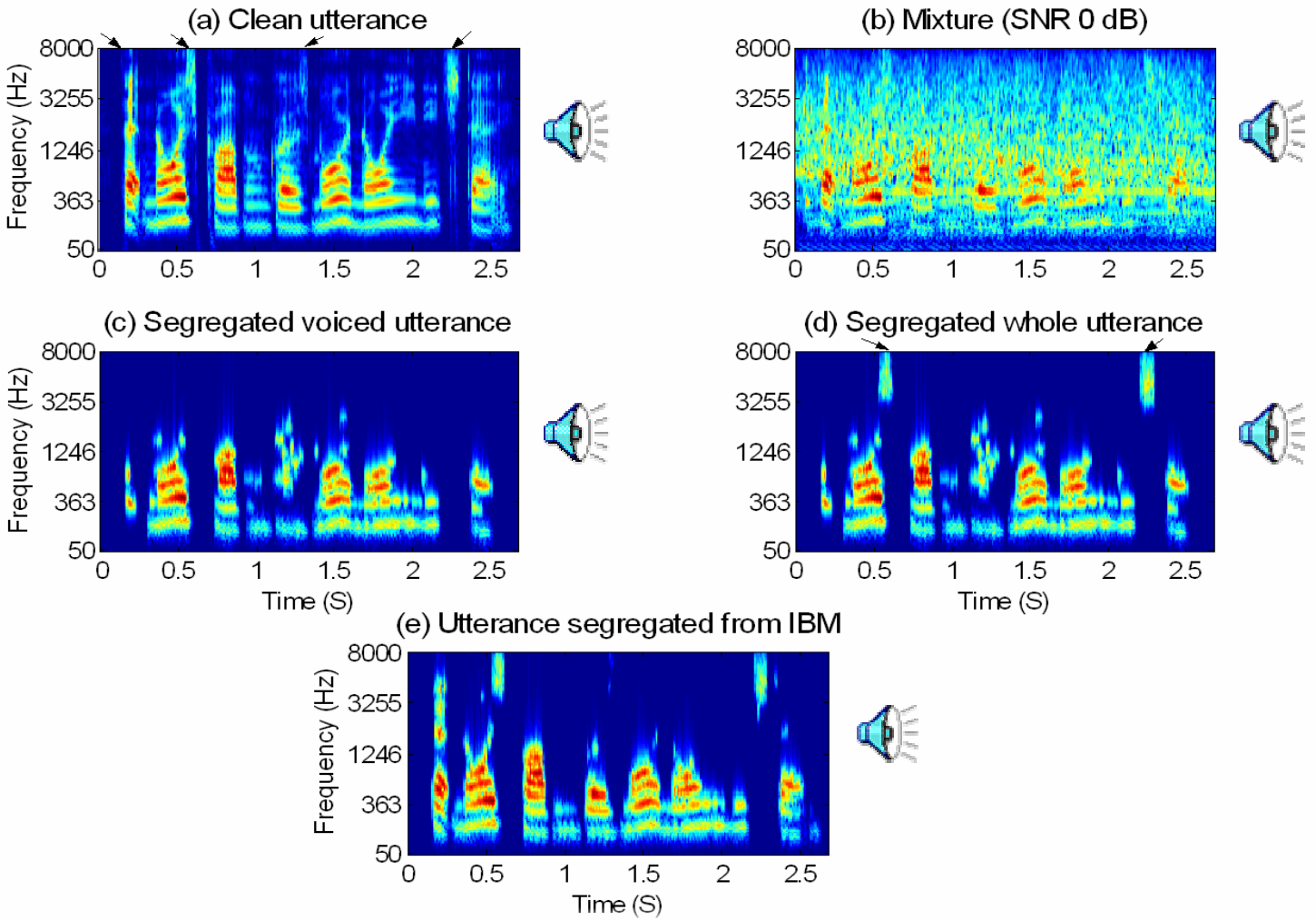
Segmentation and unvoiced speech separation

- **To deal with unvoiced speech segregation, Hu and Wang (2004) recently proposed a model of auditory segmentation that applies to both voiced and unvoiced speech**
- **The task of segmentation is to decompose an auditory scene into contiguous T-F regions, each of which should contain signal from the same sound source**
  - The definition of segmentation does not distinguish between voiced and unvoiced sounds
- **This is equivalent to identifying onsets and offsets of individual T-F regions, which generally correspond to sudden changes of acoustic energy**
- **The segmentation strategy is based on onset and offset analysis**

Scale-space analysis for auditory segmentation

- **From a computational standpoint, auditory segmentation is similar to image (visual) segmentation**
  - Visual segmentation: Finding bounding contours of visual objects
  - Auditory segmentation: Finding onset and offset fronts of segments
- **Onset/offset analysis employs scale-space theory, which is a multiscale analysis commonly used in image segmentation**
  - Smoothing
  - Onset/offset detection and onset/offset front matching
  - Multiscale integration

Example for segregating fricatives/affricates



Utterance: “That noise problem grows more annoying each day”
Interference: Crowd noise with music (IBM: Ideal binary mask)

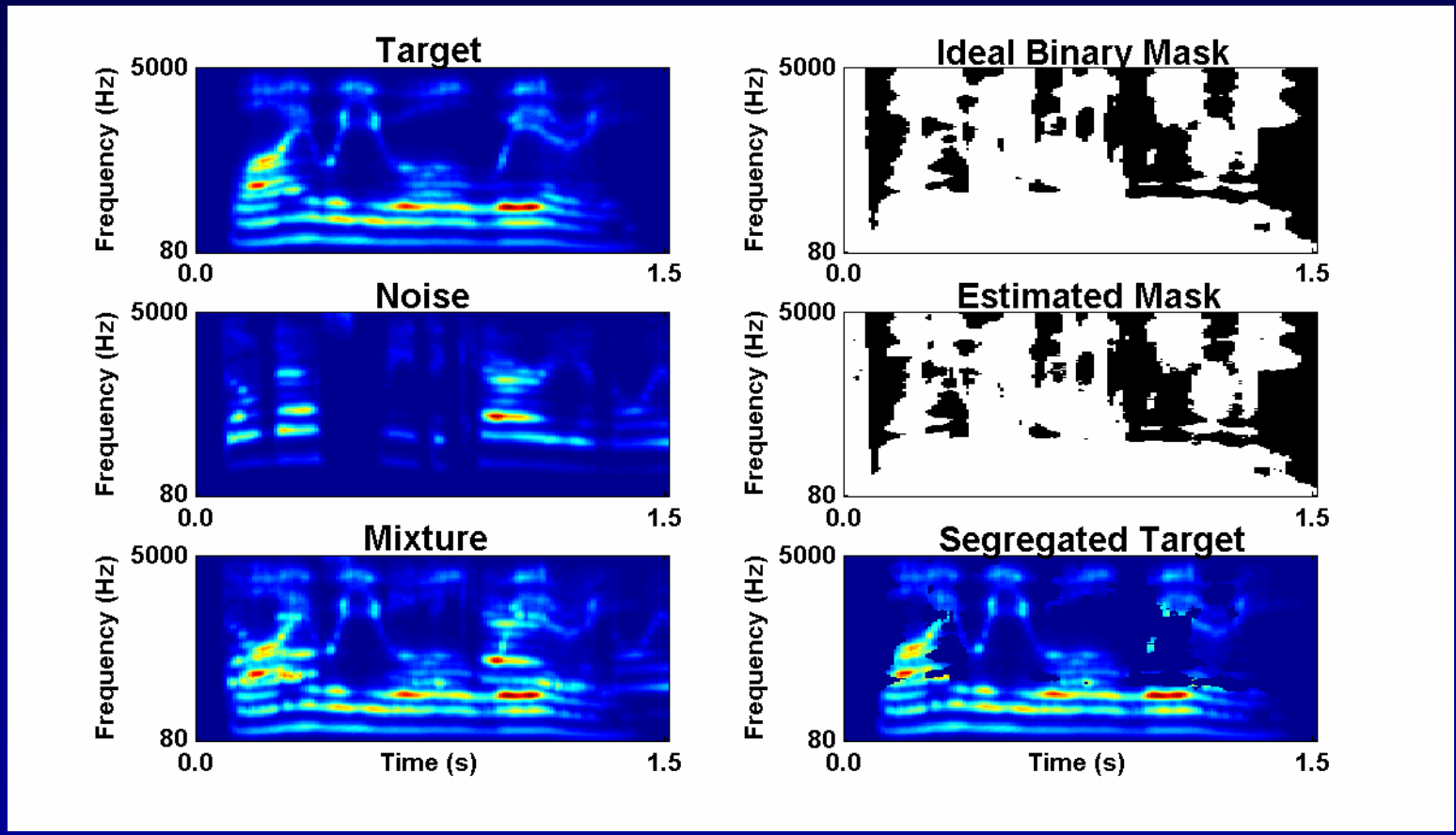
Binaural segregation of natural speech

- **Binaural speech segregation is applicable to both voiced and unvoiced speech**
- **The binaural segregation model of Roman, Wang, & Brown (2003) focuses on localization cues:**
  - Interaural time difference (ITD)
  - Interaural intensity difference (IID)
- **The model explicitly estimates ideal binary masks using supervised learning**

Ideal binary mask estimation

- **For narrowband stimuli, we observe that systematic changes of extracted ITD and IID values occur as the relative strength of the target signal (vs. the mixture) changes. This interaction produces characteristic clustering in the joint ITD-IID space**
- **The core of the model lies in deriving the statistical relation between the relative strength and the binaural cues**
  - Independent supervised training for different spatial configurations and different frequency bands in the joint ITD-IID space

Example (target: 0°, noise: 30°)



Target Noise Mixture Ideal binary mask Result

Summary and discussion

- **It pays to have an additional microphone**
  - Binaural segregation produces better results than monaural segregation
  - It works equally well for voiced and unvoiced speech
- **Binaural segregation employs spatial cues, whereas monaural segregation exploits intrinsic sound characteristics**
- **Limitations of binaural (and microphone array) segregation**
  - Cannot deal with single-microphone mixtures
  - *Configuration stationarity*: What if the target sound switches between different sound sources, or the target changes its location and orientation?
- **Can one achieve general separation without analyzing sound characteristics?**