Auditory Segmentation and Unvoiced Speech Segregation

DeLiang Wang & Guoning Hu

Perception & Neurodynamics Lab

The Ohio State University

Outline of presentation

- **Introduction**
  - Auditory scene analysis
  - Unvoiced speech problem
- **Auditory segmentation based on event detection**
- **Unvoiced speech segregation**
- **Summary**

Speech segregation

- **In a natural environment, speech is usually corrupted by acoustic interference. Speech segregation is critical for many applications, such as automatic speech recognition and hearing prosthesis**
- **Most speech separation techniques, e.g. beamforming and blind source separation via independent analysis, require multiple sensors. However, such techniques have clear limits**
  - Suffer from configuration stationarity
  - Can’t deal with single-microphone mixtures
- **Most speech enhancement developed for monaural situation can deal with only stationary acoustic interference**

Auditory scene analysis (ASA)

- **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 'sensory elements' (segments)
  - **Grouping**. Combine segments into streams so that the segments of the same stream likely originate from the same source

Computational auditory scene analysis

- **Computational ASA (CASA) approaches sound separation based on ASA principles**
- **CASA successes: Monaural segregation of voiced speech**
- **A main challenge is segregation of unvoiced speech, which lacks the periodicity cue**

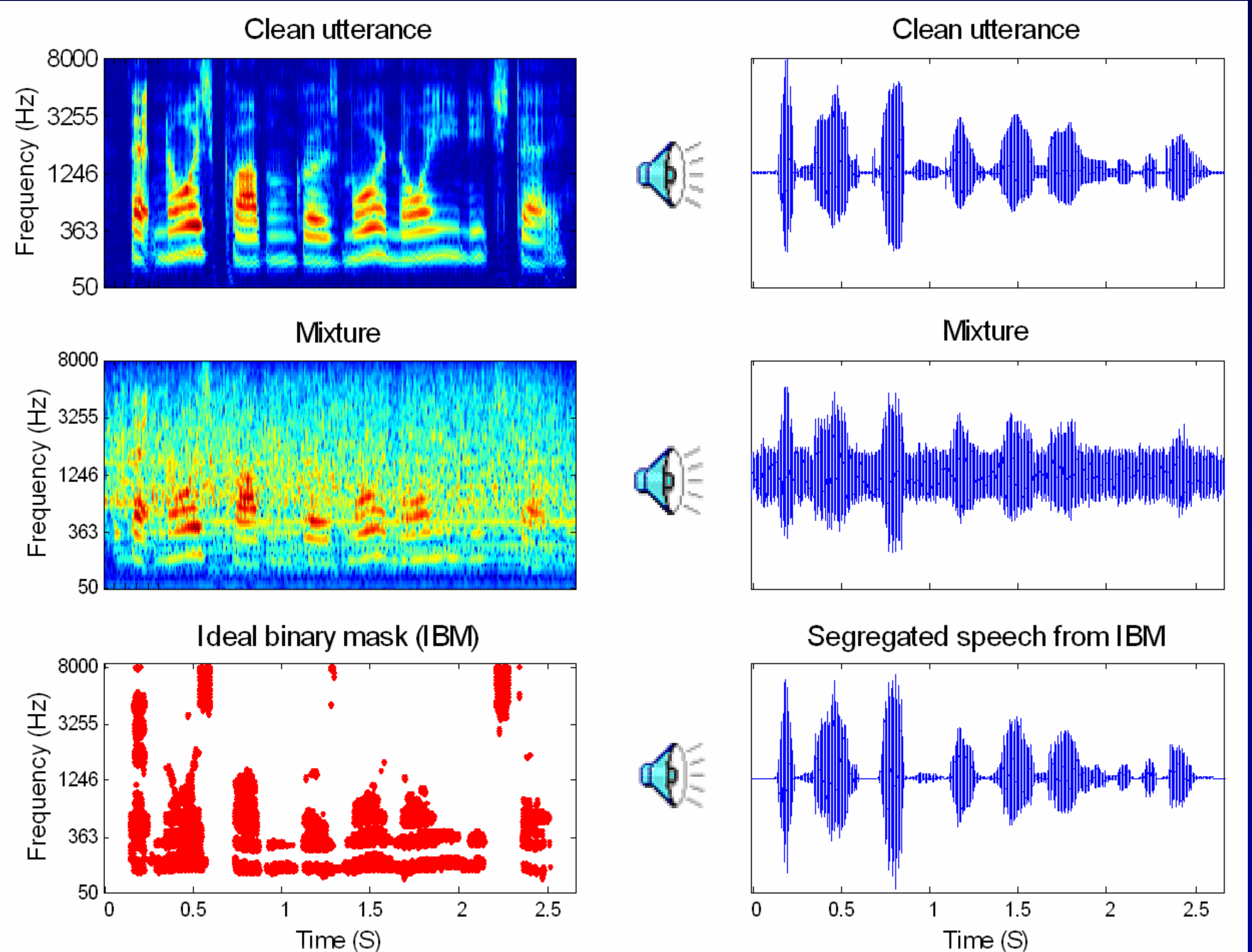
Unvoiced speech

- **Speech sounds consist of vowels and consonants, the latter are further composed of voiced and unvoiced consonants**
- **For English, the relative frequencies of different phoneme categories are (Dewey'23):**
  - Vowels: 37.9%
  - Voiced consonants: 40.3%
  - Unvoiced consonants: 21.8%
- **In terms of time duration, unvoiced consonants account for about 1/5 in American English**
- **Consonants are crucial for speech recognition**

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 SNR criterion for mask generation

Ideal binary masking illustration



Utterance: “That noise problem grows more annoying each day”
Interference: Crowd noise with music (0 SNR)

Outline of presentation

- **Introduction**
  - Auditory scene analysis
  - Unvoiced speech problem
- **Auditory segmentation based on event detection**
- **Unvoiced speech segregation**
- **Summary**

Auditory segmentation

- **Our approach to unvoiced speech segregation breaks the problem into two stages: segmentation and grouping**
  - This presentation is mainly about segmentation
- **The task of segmentation is to decompose an auditory scene into contiguous T-F regions, each of which should contain signal from the same event**
  - It should work for both 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**
- **Our segmentation strategy is based on onset and offset analysis of auditory events**

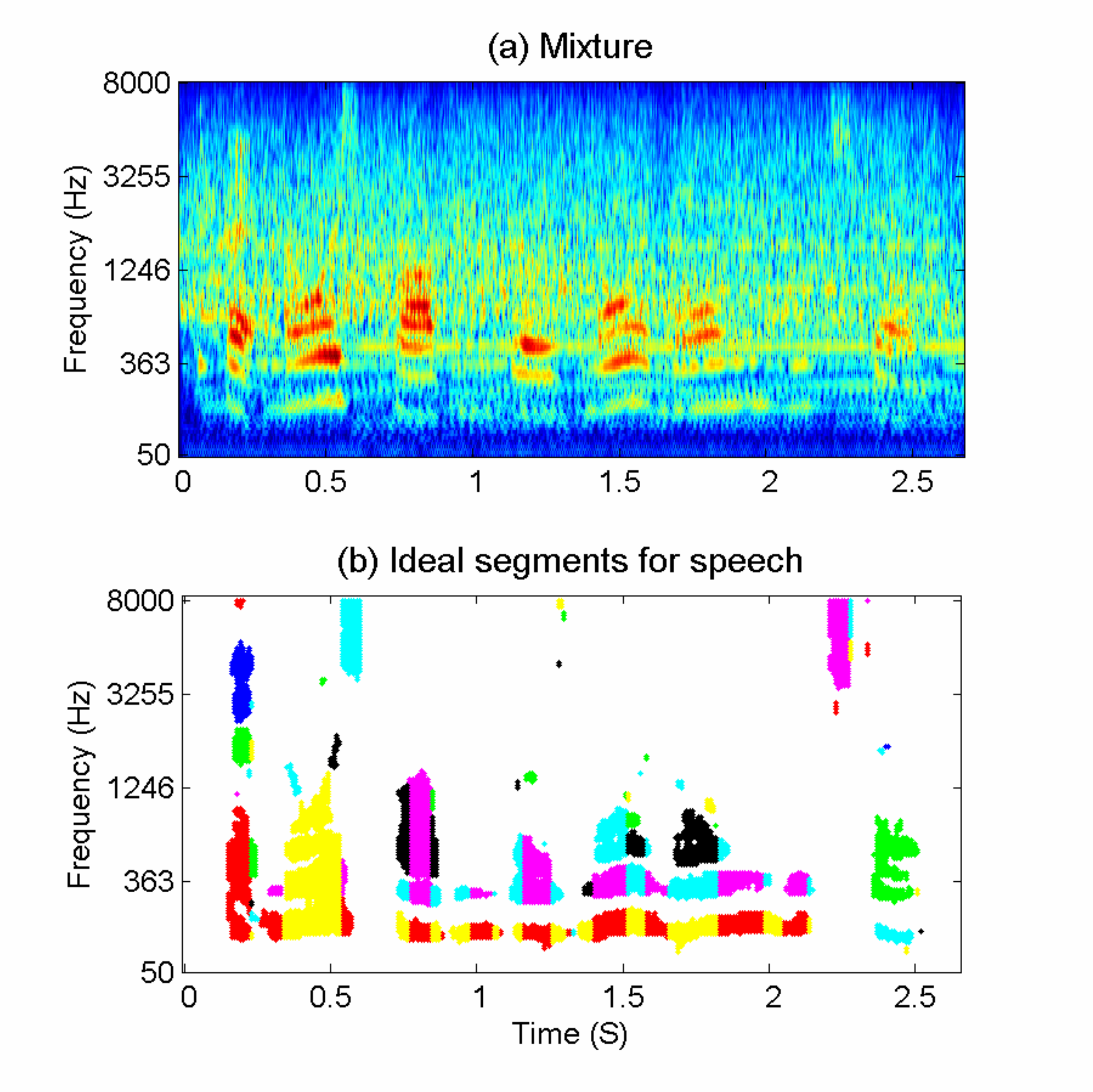
What is an auditory event?

- **To define an auditory event, two perceptual effects need to be considered:**
  - Audibility
  - Auditory masking
- **We define an auditory event as a collection of the audible T-F regions from the same sound source that are stronger than combined intrusions**
- **Hence the computational goal of segmentation is to produce segments, or contiguous T-F regions, of an auditory event**
  - For speech, a segment corresponds to a phone

Cochleogram as a peripheral representation

- **We decompose an acoustic input using a gammatone filterbank**
  - 128 filters centered from 50 Hz to 8 kHz
  - Filtering is performed in 20-ms time frames with 10-ms frame shift
- **The intensity output forms what we call a cochleogram**

Cochleogram and ideal segments



Scale-space analysis for auditory segmentation

- **From a computational standpoint, auditory segmentation is similar to image segmentation**
  - Image segmentation: Finding bounding contours of visual objects
  - Auditory segmentation: Finding onset and offset fronts of segments
- **Our onset/offset analysis employs scale-space theory, which is a multiscale analysis commonly used in image segmentation**
- **Our proposed system performs the following computations:**
  - Smoothing
  - Onset/offset detection and matching
  - Multiscale integration

Smoothing

- **For each filter channel, the intensity is smoothed over time to reduce the intensity fluctuation**
- **An event tends to have onset and offset synchrony in the frequency domain. Consequently the intensity is further smoothed over frequency to enhance common onsets and offsets in adjacent frequency channels**
- **Smoothing is done via dynamic diffusion**

Smoothing via diffusion

- **A one-dimensional diffusion of a quantity *v* across the spatial dimension *x* is governed by:**

$$\frac{\partial v}{\partial t} = \frac{\partial}{\partial x}\left[D(v) \cdot \frac{\partial v}{\partial x} \right]$$

- ***D* is a function controlling the diffusion process. As *t* increases, *v* gradually smoothes over *x***
- **The diffusion time *t* is called the scale parameter and the smoothed *v* values at different times compose a scale space**

Diffusion

- **Let the input intensity be the initial value of *v*, and let *v* diffuse across time frames, *m*, and filter channels, *c*, as follows:**

$$v(c,m,0,0) = I(c,m)$$

$$\frac{\partial v(c,m,0,t_m)}{\partial t_c} = \frac{\partial}{\partial m}[D_m(v) \cdot \frac{\partial v}{\partial m}]$$

$$\frac{\partial v(c,m,t_c,t_m)}{\partial t_c} = \frac{\partial}{\partial c}[D_c(v) \cdot \frac{\partial v}{\partial c}]$$

  - $I(c, m)$ is the logarithmic intensity in channel c at frame m

Diffusion, continued

- **Two forms of $D_m(v)$ are employed in the time domain:**
  - $D_m(v) = 1$, which reduces to Gaussian smoothing:

$$v(c,m,0,t_m) = v(c,m,0,0) * G(0,2t_m)$$

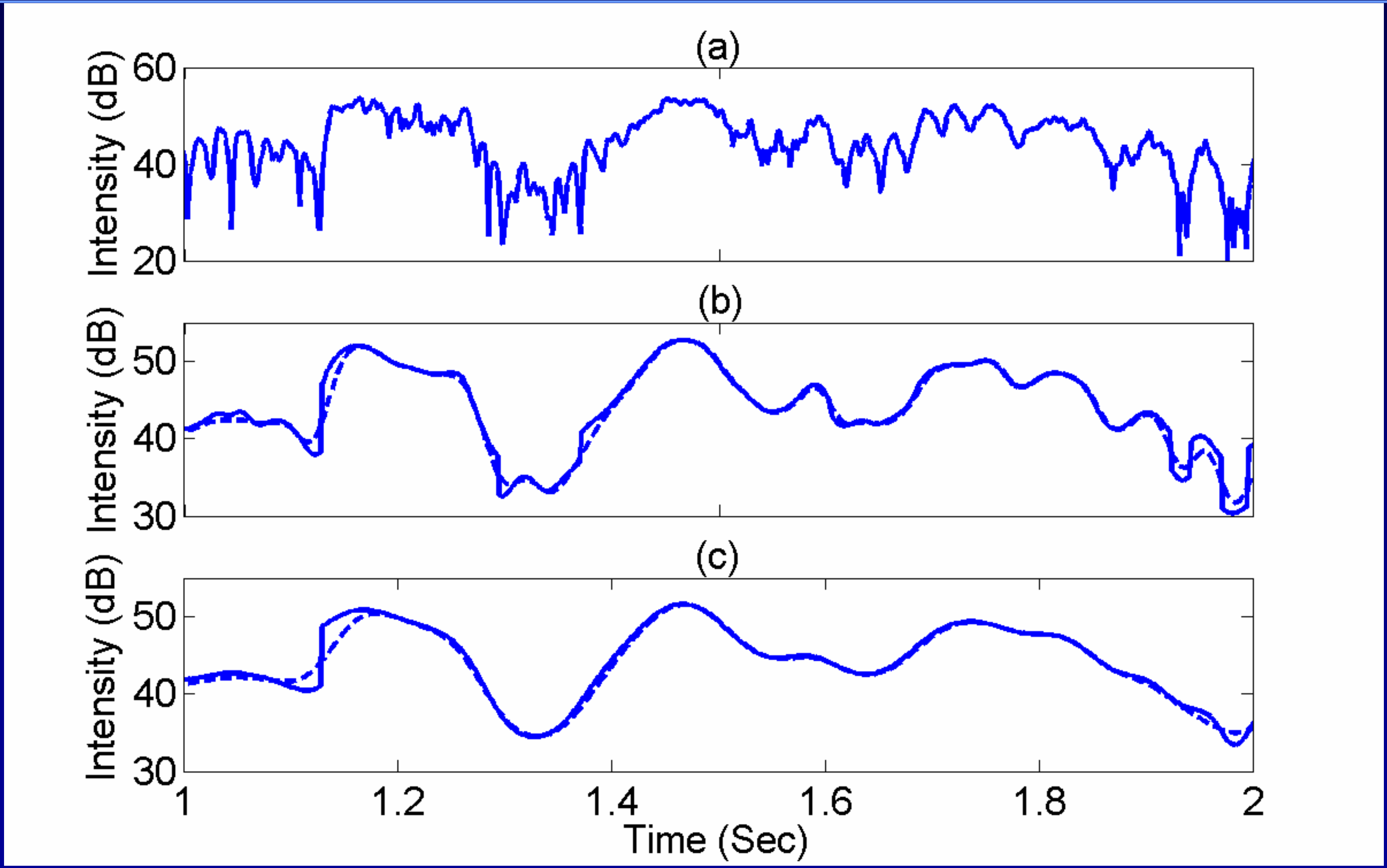
  - Perona-Malik ('90) anisotropic diffusion:

$$D_m(v) = 1/[1 + |\frac{\partial v}{\partial m}|^2 / \lambda^2]$$

    Compared with Gaussian smoothing, the Perona-Malik model may identify onset and offset positions bettter

- **In the frequency domain, $D_c(v) = 1$**

Diffusion results



Top: Initial intensity. Middle and bottom: Two scales for Gaussian smoothing (dash line) and anisotropic diffusion (solid line)

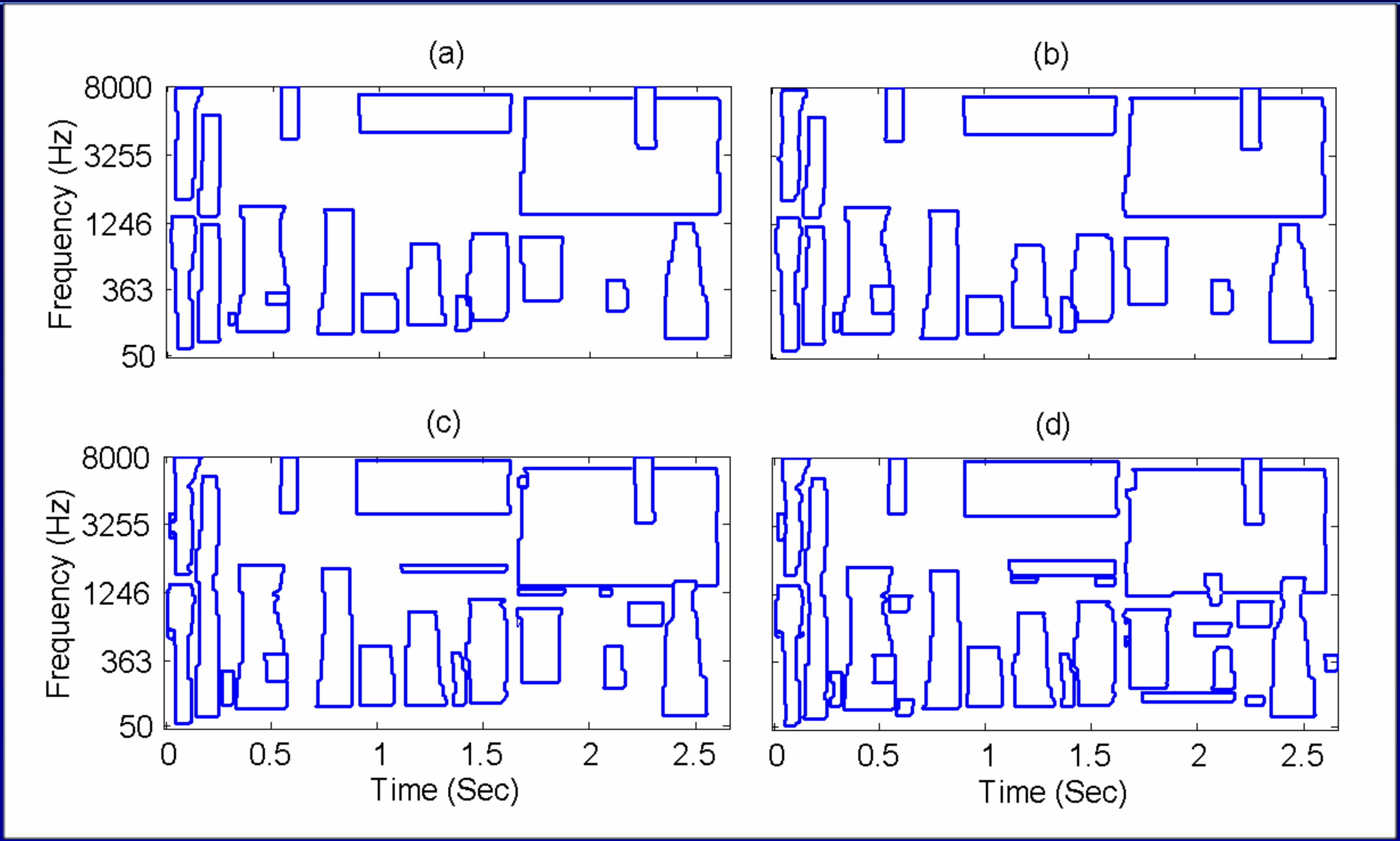
Onset/offset detection and matching

- **At each scale, onset and offset candidates are detected by identifying peaks and valleys of the first-order time-derivative of *v***
- **Detected candidates are combined into onset and offset fronts, which form vertical curves**
- **Individual onset and offset fronts are matched to yield segments**

Multiscale integration

- **The system integrates segments generated with different scales iteratively:**
  - First, it produces segments at a coarse scale (more smoothing)
  - Then, at a finer scale, it locates more accurate onset and offset positions for these segments. In addition, new segments may be produced
- **The advantage of multiscale integration is that it analyzes an auditory scene at different levels of detail so as to detect and localize auditory segments at appropriate scales**

Segmentation at different scales



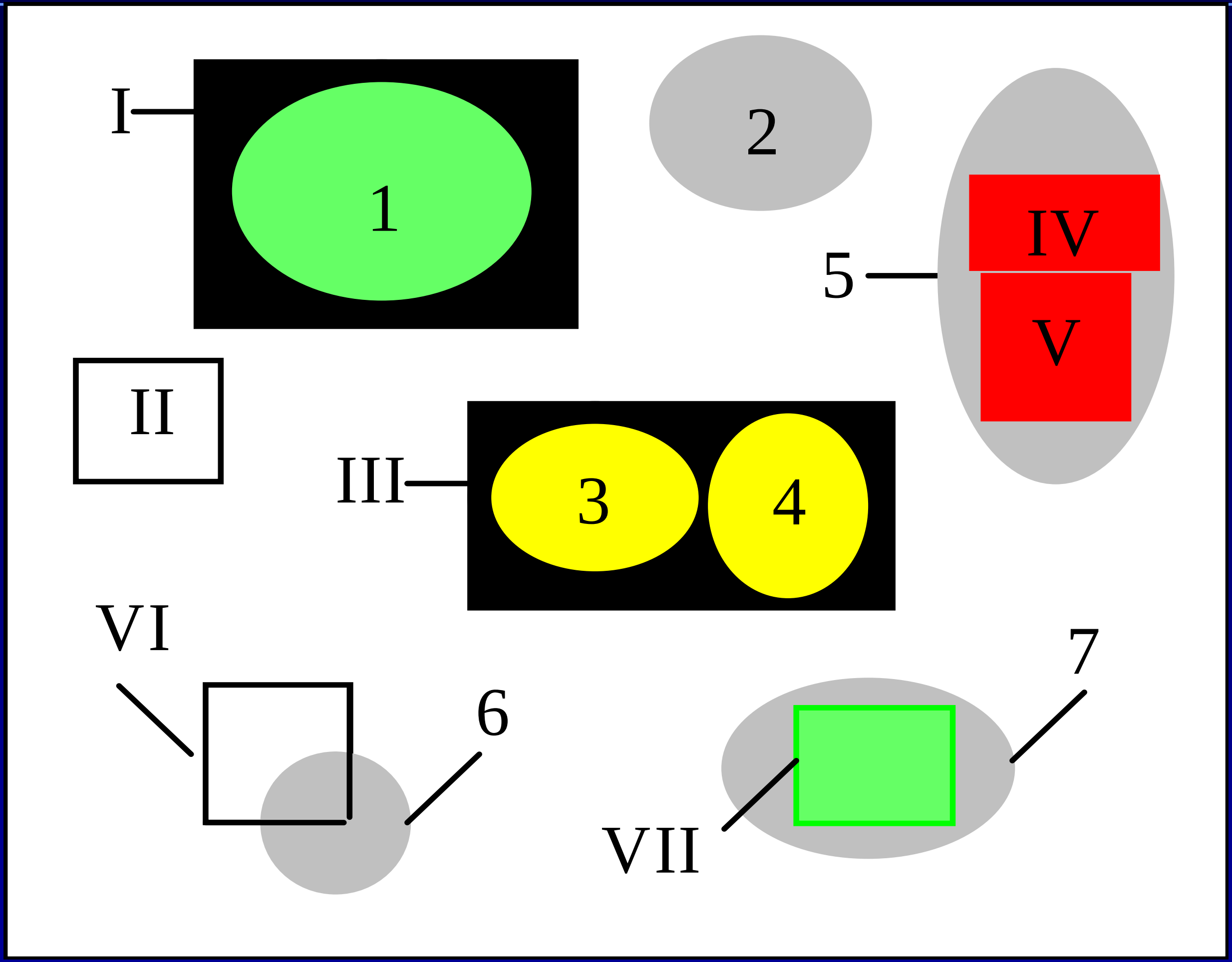
Input: Mixture of speech and crowd noise with music

Scales (t_c, t_m) are: (a). (32, 200); (b). (18, 200); (c). (32, 100). (d). (18, 100)

Evaluation

- **How to quantitatively evaluate segmentation results is a complex issue, since one has to consider various types of mismatch between a collection of ideal segments and that of computed segments**
- **Here we adapt a region-based definition by Hoover *et al.* ('96), originally proposed for evaluating image segmentation systems**
- **Based on the degree of overlapping (defined by threshold θ), we label a T-F region as belonging to one of the five classes**
  - Correct
  - Under-segmented. Under-segmentation is not really an error because it produces larger segments – good for subsequent grouping
  - Over-segmented
  - Missing
  - Mismatching

Illustration of different classes



Ovals (Arabic numerals) indicate ideal segments and rectangles (Roman numerals) computed segments. Different colors indicate different classes

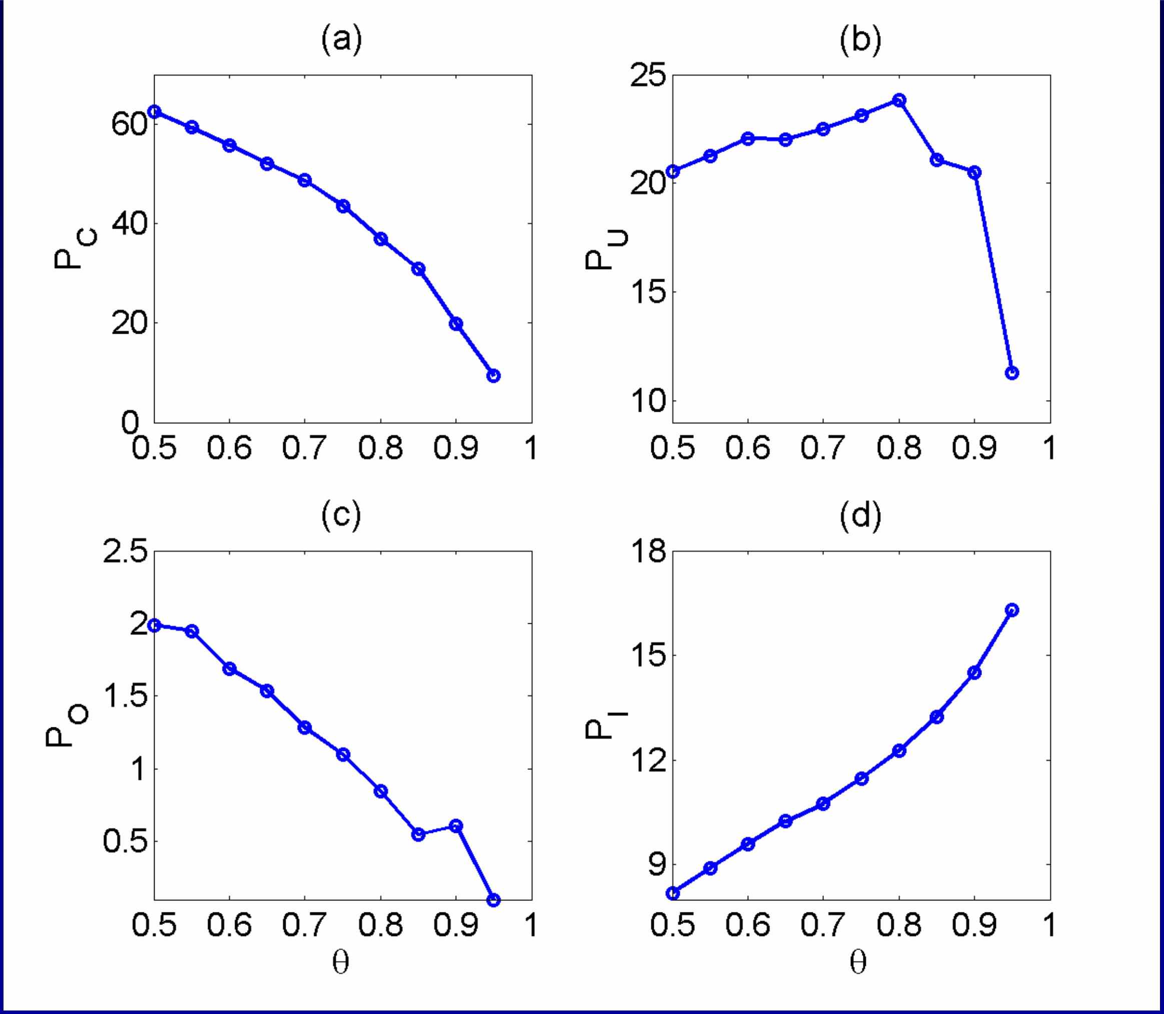
Quantitative measures

- **Let E_C, E_U, E_O, E_M, and E_I be the summated energy in all the regions labeled as correct, under-segmented, over-segmented, missing, and mismatching respectively. Let E_{GT} be the total energy of all ideal segments and E_S that of all estimated segments**
  - The percentage of correctness: $P_C = E_C / E_{GT} \times 100\%$.
  - The percentage of under-segmentation: $P_U = E_U / E_{GT} \times 100\%$.
  - The percentage of over-segmentation: $P_O = E_O / E_{GT} \times 100\%$.
  - The percentage of mismatch, $P_I = E_I / E_S \times 100\%$.
  - The percentage of missing, $P_M = (1 - P_C - P_U - P_O) \times 100\%$.

Evaluation corpus

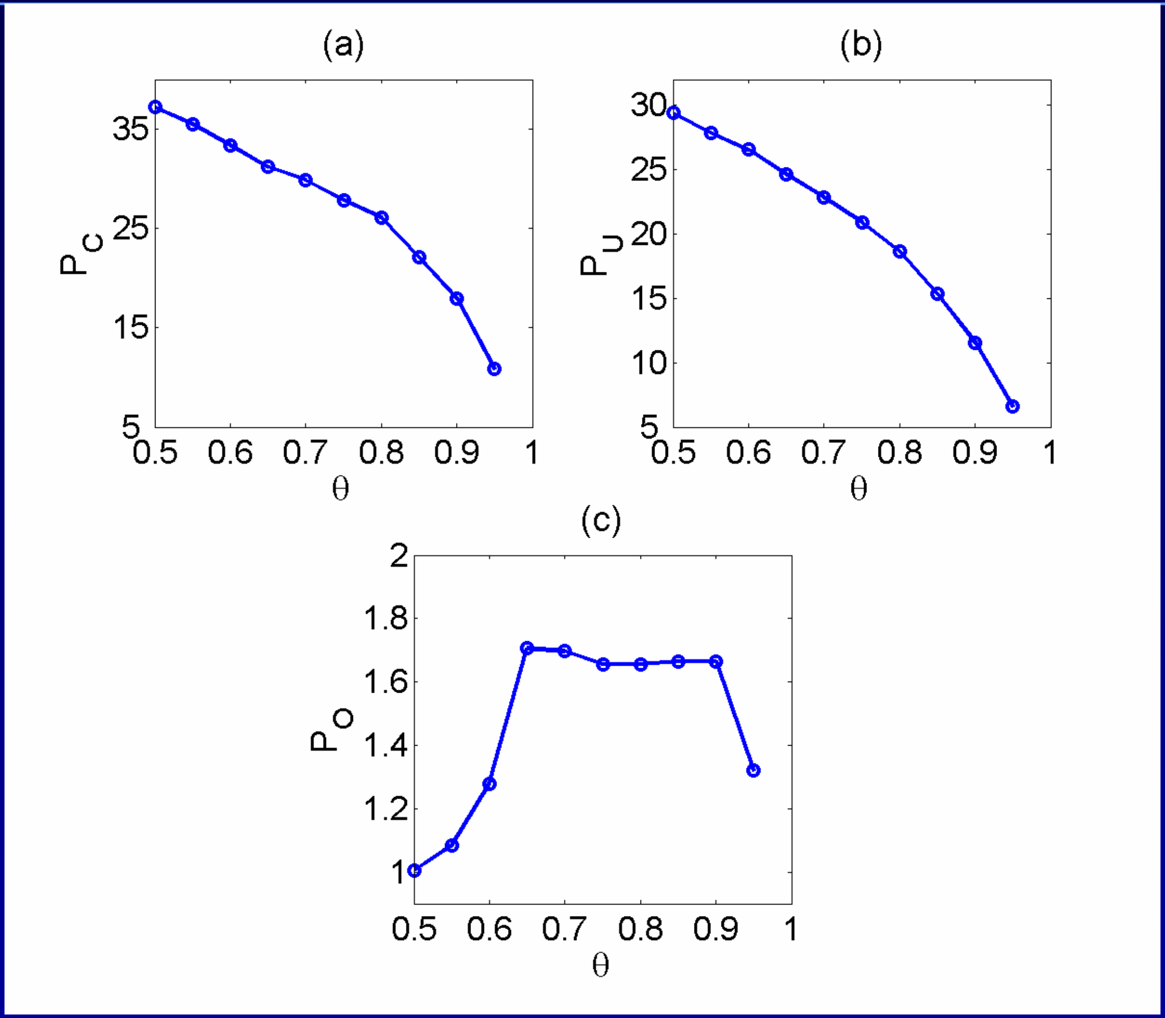
- **20 utterances from the TIMIT database**
- **10 types of intrusion: white noise, electrical fan, rooster crowing and clock alarm, traffic noise, crowd in playground, crowd with music, crowd clapping, bird chirping and waterflow, wind, and rain**

Results on all phonemes

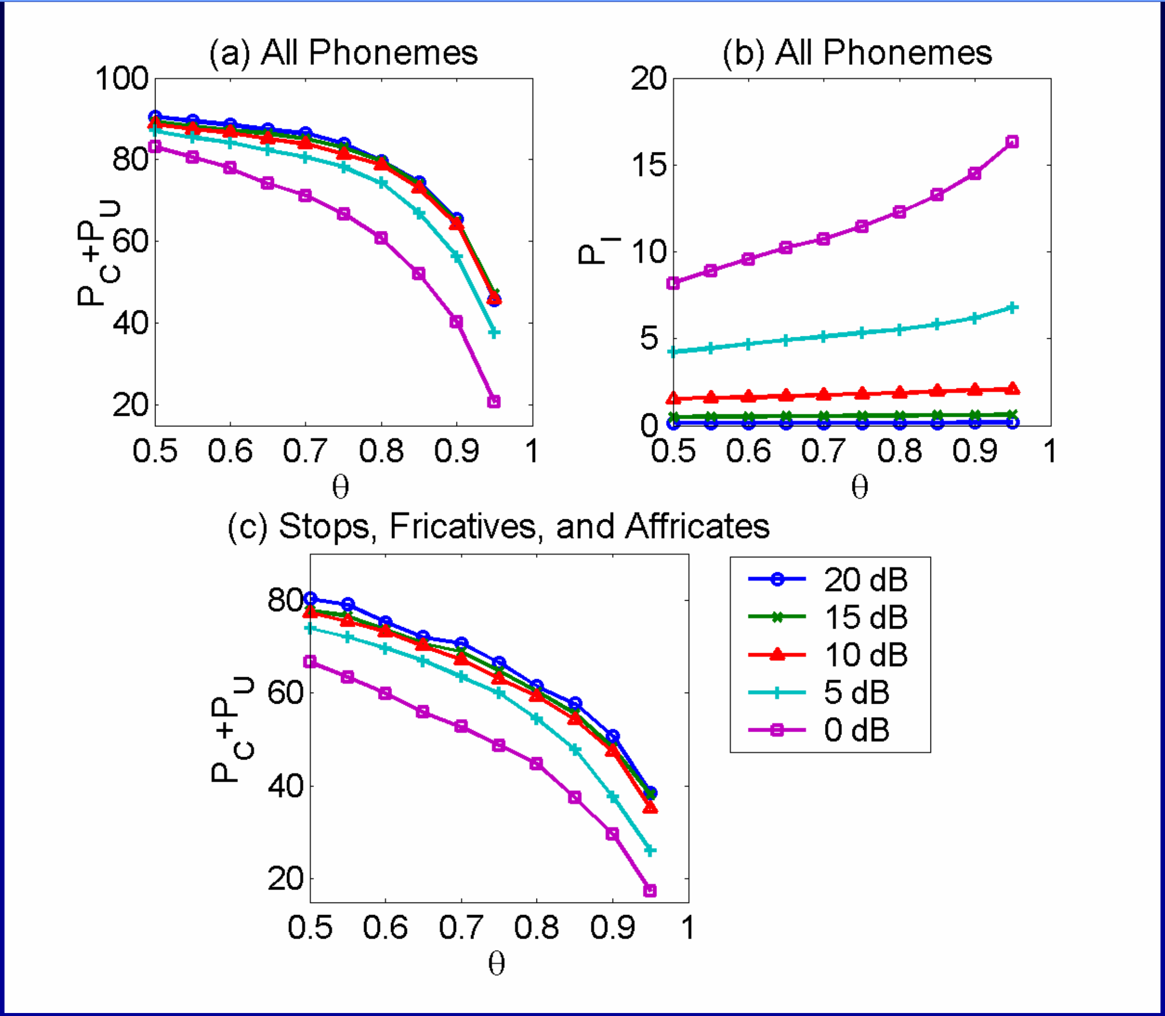


Results are with respect to θ, with 0 dB mixtures and anisotropic diffusion

Results on stops, fricatives, and affricates

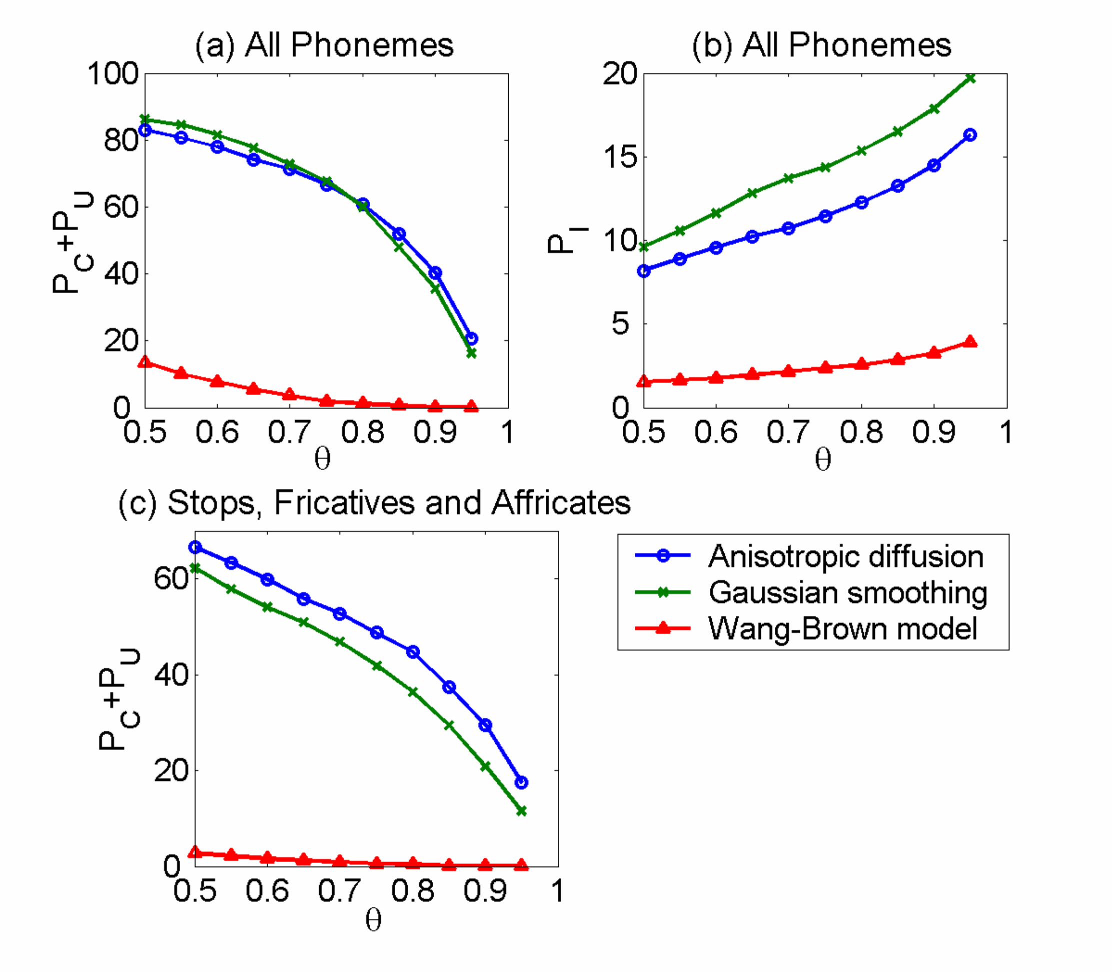


Results with different mixture SNRs



P_C and P_U are combined here since P_U is not really error

Comparisons



Comparisons are made between anisotropic diffusion and Gaussian smoothing, as well as with the Wang-Brown model (1999), which deals with mainly with voiced segments using cross-channel correlation. Mixtures are at 0 dB SNR

Outline of presentation

- **Introduction**
  - Auditory scene analysis
  - Unvoiced speech problem
- **Auditory segmentation based on event detection**
- **Unvoiced speech segregation**
- **Summary**

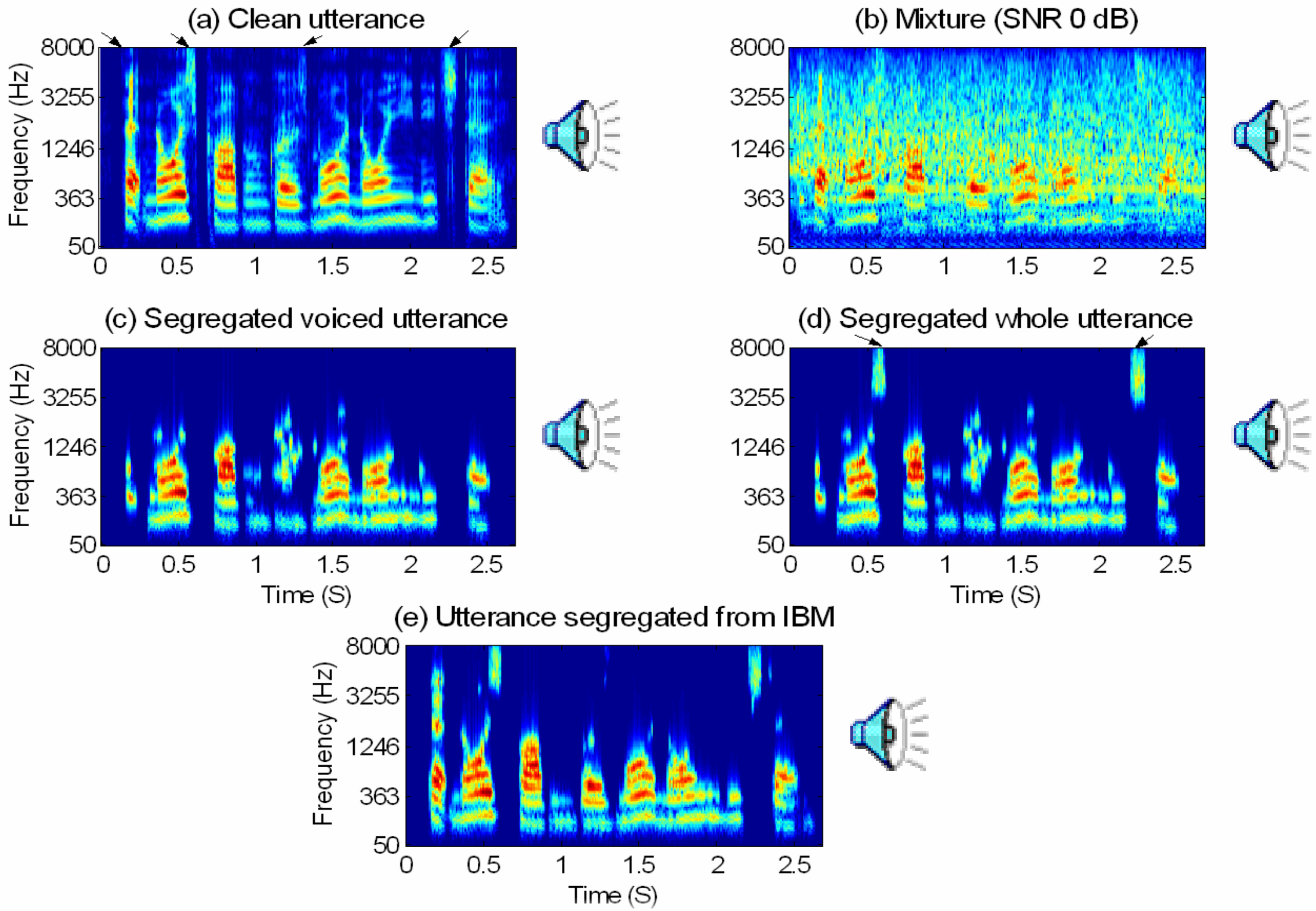
Speech segregation

- **The general strategy for speech segregation is to first segregate voiced speech using the pitch cue, and then deal with unvoiced speech**
- **Voiced speech segregation is performed using our recent model (Hu & Wang'04):**
  - The model generates segments for voiced speech using cross-channel correlation and temporal continuity
  - It groups segments according to periodicity and amplitude modulation
- **To segregate unvoiced speech, we perform auditory segmentation, and then group segments that correspond to unvoiced speech**

Segment classification

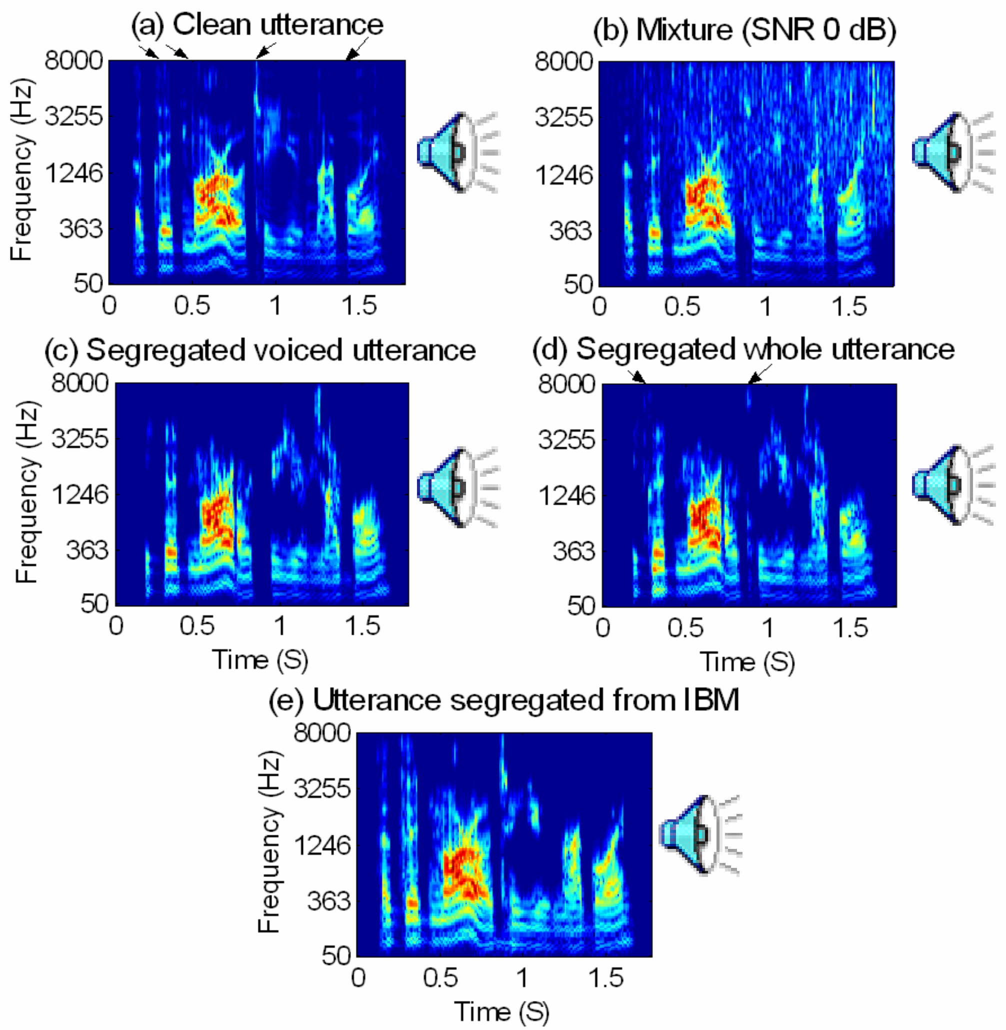
- **For nonspeech interference, grouping is in fact a classification task – to classify segments as either speech or non-speech**
- **The following features are used for classification:**
  - Spectral envelope
  - Segment duration
  - Segment intensity
- **Training data**
  - Speech: Training part of the TIMIT database
  - Interference: 90 natural intrusions including street noise, crowd noise, wind, etc.
- **A Gaussian mixture model is trained for each phoneme, and for interference as well which provides the basis for a likelihood ratio test**

Demo for fricatives and affricates



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

Demo for stops



Utterance: "A good morrow to you, my boy"
Interference: Rain

Summary

- **We have proposed a model for auditory segmentation, based on a multiscale analysis of onsets and offsets**
- **Our model segments both voiced and unvoiced speech sounds**
- **The general strategy for unvoiced (and voiced) speech segregation is to first perform segmentation and then group segments using various ASA cues**
- **Sequential organization of segments into streams is not addressed**
- **How well can people organize unvoiced speech?**