On Computational Objectives of Auditory Scene Analysis

DeLiang Wang

The Ohio State University

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
On Computational Objectives of Auditory Scene Analysis	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Ohio State University,Department of Computer Science and Engineering,and the Center of Cognitive Science,Columbus,OH,43210**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
NSF Workshop on Perceptives on Speech Separation, Montreal, Canada, Nov. 2003. U.S. Government or Federal Rights License

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**36**	

Outline of Presentation

- **Introduction**
  - Sound source separation problem
  - Approaches to sound separation
  - Auditory scene analysis (ASA)
- **Computational ASA and its objectives**
- **Ideal binary masks as a putative objective**
- **Example studies of computing ideal binary masks**
  - Monaural segregation of voiced speech
  - Binaural segregation of natural speech
- **Summary**

Sound Source Separation Problem

- **In a natural environment, a target sound source (e.g. speech) is usually accompanied by acoustic interference** 
- **Many sound processing tasks, such as automatic speech recognition, audio retrieval, and hearing aid design, require a solution to the sound separation problem**
- **Problem has been studied using different approaches**

Approaches to Sound Separation Problem

- **Speech enhancement: Enhance signal-to-noise ratio (SNR) or speech quality by attenuating interference**
  - Advantage: Simple and applicable to one-microphone recordings
  - Challenge: Prior knowledge of interference
- **Spatial filtering (beamforming): Extract target sound from a specific spatial direction with a sensor array**
  - Advantage: High fidelity and robustness to reverberation
  - Challenge: Rigidity. What if target switches or changes its location?
- **Independent component analysis: Find a demixing matrix from mixtures of sound sources**
  - Advantage: High fidelity when assumptions are met
  - Challenge: Limiting assumptions. Chief among them is stationarity of mixing matrix

Auditory Scene Analysis (Bregman'90)

- **Listeners are able to parse a complex mixture of sounds arriving at the ears in order to retrieve a mental representation of each sound source**
  - Ball-room problem, Helmholtz, 1863 ("complicated beyond conception")
  - Cocktail-party problem, Cherry'53
- **Two conceptual processes of ASA:**
  - **Segmentation**. Decompose the acoustic mixture into sensory elements (segments)
  - **Grouping**. Combine segments into groups, so that segments in the same group are likely to have originated from the same source

Computational Auditory Scene Analysis

- **Computational ASA (CASA) approaches sound separation based on ASA principles**
  - Weintraub’85, Cooke’93, Brown & Cooke’94, Klassner’96, Ellis’96, Wang & Brown’99
  - Problem domain or technical approach?
- **CASA advantage: Monaural segregation with minimal assumptions**
- **CASA challenge: Reliable pitch tracking of noisy speech, unvoiced speech, room reverberation**

CASA Evaluation Criteria

- **Comparing segregated target with premixing target**
  - In terms of the group of target elements (Cooke'93)
  - In terms of SNR (Brown & Cooke'94; Wang & Brown'99)
  - In terms of spectral distortion (Nakatani & Okuno'99) or Wiener filter (Bodden'93)
- **Automatic speech recognition (ASR)**
  - Weintraub'85; Glottin'01
- **Human listening**
  - Stubbs and Summerfield'90; Ellis'96
- **Fit with perceptual and biological phenomena**
  - Wang'96; McCabe and Denham'97; Wrigley'02

What Is the Goal of CASA?

- **What is the goal of perception?**
  - The perceptual systems are ways of seeking and extracting information about the environment from sensory input (Gibson'66)
  - The purpose of vision is to produce a visual description of the environment for the viewer (Marr'82)
  - By analogy, the purpose of audition is to produce an auditory description of the environment for the listener
- **What is the computational goal of ASA?**
  - The goal of ASA is to segregate sound mixtures into separate perceptual representations (or auditory streams), each of which corresponds to an acoustic event (Bregman'90)
  - By extrapolation the goal of CASA is to develop computational systems that extract individual streams from sound mixtures

Marrian Three Levels of Analysis

- **According to Marr (1982), a complex information processing system must be understood in three levels**
  - Computational theory: goal, its appropriateness, and basic processing strategy
  - Representation and algorithm: representations of input and output and transformation algorithms
  - Implementation: physical realization
- **All levels of explanation are required for eventual understanding of perceptual information processing**
- **Computational theory analysis – understanding the character of the problem – is critically important**

Computational-Theory Analysis of ASA

- **To form a stream, a sound must be audible on its own**
- **The number of streams that can be computed at a time is limited**
  - Magical number 4 for simple sounds such as tones and vowels (Cowan'01)?
  - 1, or figure-ground segregation, in noisy environment such as a cocktail party?
- **Auditory masking further constrains the ASA output**
  - Within a critical band a stronger signal masks a weaker one 

Computational-theory Analysis of ASA - continued

- **ASA result depends on sound types (overall SNR is 0)**
  - Noise-Noise: pink, white, pink+white
  - Tone-Tone: tone1, tone2 , tone1+tone2
  - Speech-Speech:
  - Noise-Tone:
  - Noise-Speech:
  - Tone-Speech:

Some Alternative CASA Objectives

- **Extract all underlying sound sources or a target sound source**
  - Segregating all sources is implausible (probably unrealistic with one or two microphones)
  - A target might be too soft to be segregated
- **Enhance ASR**
  - Advantage: close coupling with a primary motivation of CASA
  - Disadvantage
    - Specific to one kind of signal (e.g. what about music?)
    - Perceiving is more than recognizing (Treisman'99)
- **Enhance human listening**
  - Advantage: close coupling with auditory perception
  - Disadvantage
    - There are CASA applications that involve no human listening
    - Not always feasible for engineers

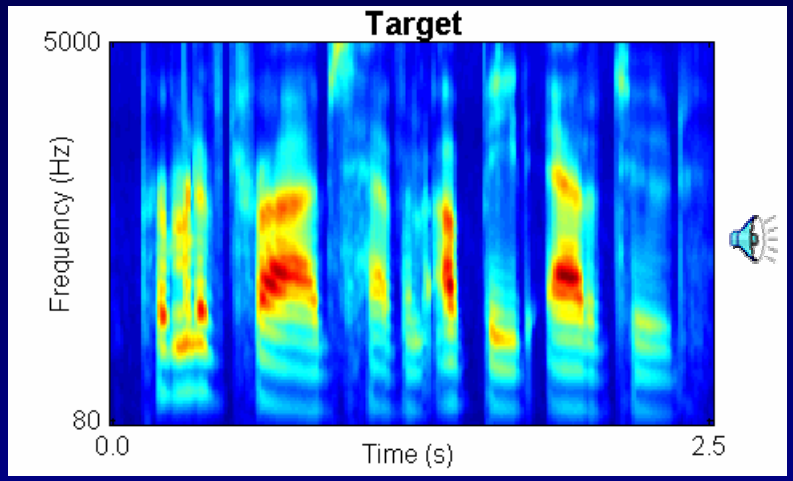
Outline of Presentation

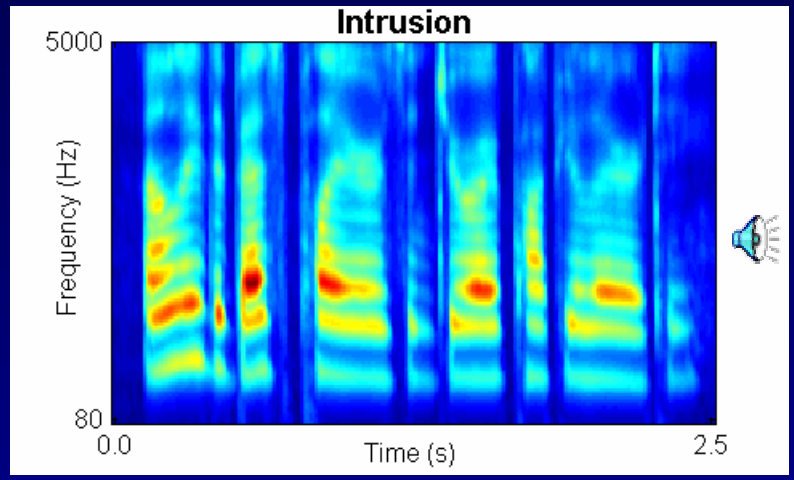
- **Introduction**
  - Sound source separation problem
  - Approaches to sound separation
  - Auditory scene analysis (ASA)
- **Computational ASA and its objectives**
- **Ideal binary masks as a putative objective**
- **Example studies of computing ideal binary masks**
  - Monaural segregation of voiced speech
  - Binaural segregation of natural speech
- **Summary**

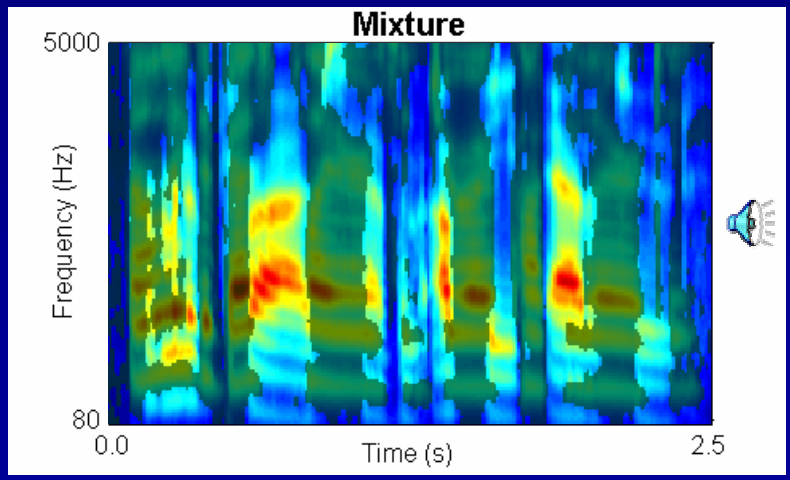
Ideal Binary Mask as a Putative Goal of CASA

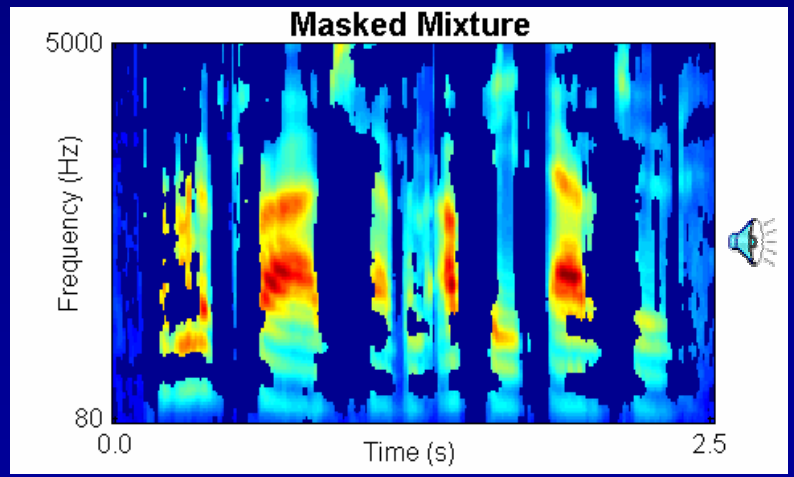
- **Key idea is to retain parts of a target sound that are stronger than the acoustic background, or to mask interference by the target**
  - What a target is depends on intention, attention, etc.
- **Within a local time-frequency (T-F) unit, the ideal binary mask is 1 if target energy is stronger than interference energy, and 0 otherwise (Hu & Wang'01; Roman et al.'03)**
  - Local 0 SNR criterion for mask generation
  - Earlier studies use binary masks as an output representation (Brown & Cooke'94; Wang and Brown'99; Roweis'00), but do not suggest the explicit notion of an ideal binary mask

Ideal Binary Mask Illustration









Resemblance to Visual Occlusion

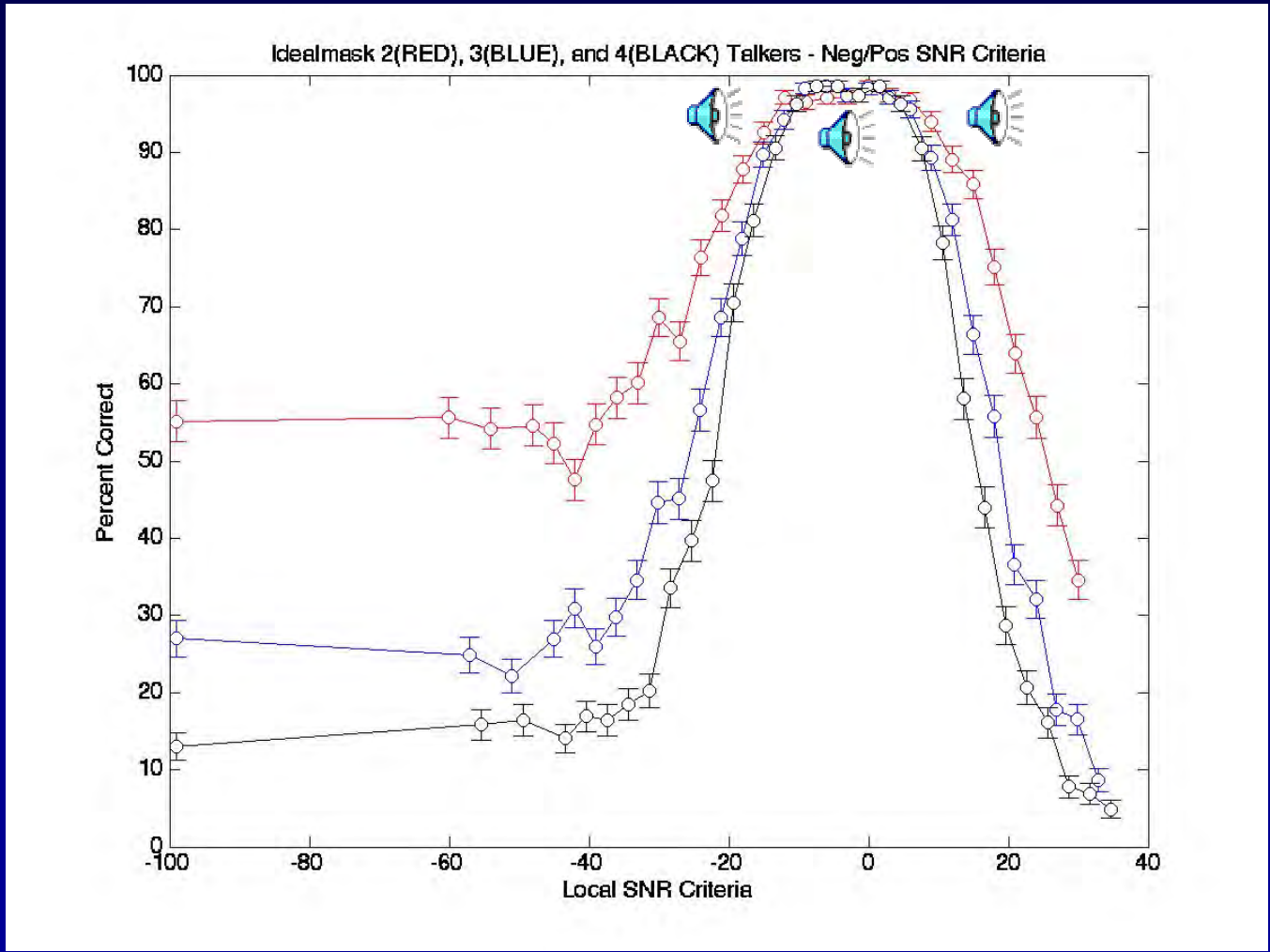
Properties of Ideal Binary Masks

- **Flexibility: With the same mixture, the definition leads to different masks depending on what target is**
- **Well-definedness: An ideal mask is well-defined no matter how many intrusions are in the scene or how many targets need to be segregated**
- **Consistent with computational-theory analysis of ASA**
  - Audibility and capacity
  - Auditory masking
- **Ideal binary masks yield good target resynthesis and provide a highly effective front-end for automatic speech recognition (Cooke et al.'01)**
  - ASR performance degrades gradually with deviations from an ideal mask (Roman et al.'03)

Ideal Binary Masking and Speech Intelligibility

- **Ideal binary masking provides a potential methodology to remove informational masking (distraction from perceptually similar maskers) by making maskers inaudible**
- **Human speech intelligibility tests on ideal binary masking (Chang, Brungart, et al.'03)**
  - Stimuli: CRM (coordinate response measure) corpus
  - 1-3 speech maskers (competing talkers)
  - Varying SNR criterion for each T-F unit

Intelligibility Results



Overall target to single-masker SNR is 0 dB

Results and Implications

- **Intelligibility performance reaches near 100% for a range of local SNR criteria, from around -10 dB to +10 dB**
  - Precise criterion for local SNR is not necessary in order to produce high intelligibility
- **Systematic degradation towards higher or lower local SNR criteria and more talkers**
- **Informational masking is eliminated**
  - Is informational masking localized energetic masking?

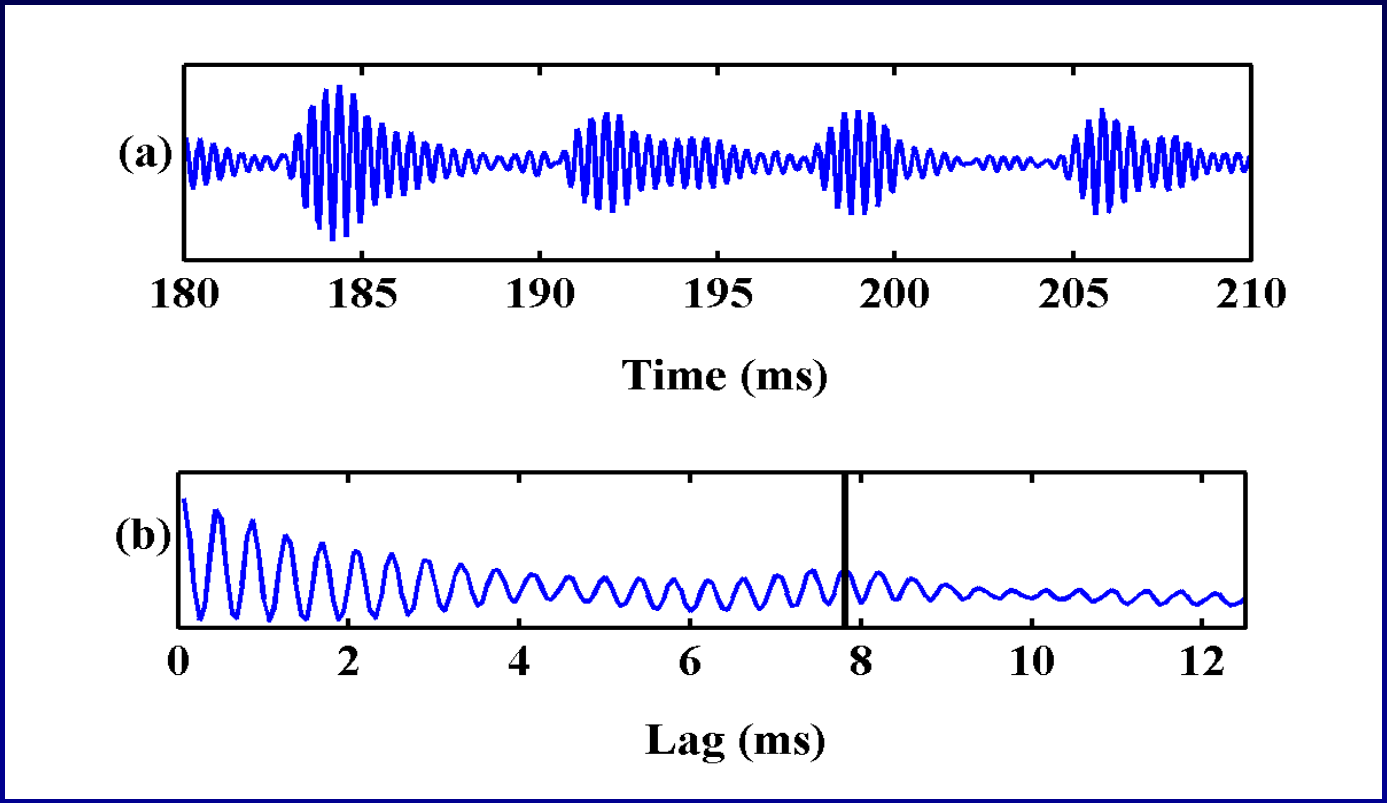
Outline of Presentation

- **Introduction**
  - Sound source separation problem
  - Approaches to sound separation
  - Auditory scene analysis (ASA)
- **Computational ASA and its objectives**
- **Ideal binary masks as a putative objective**
- **Example studies of computing ideal binary masks**
  - Monaural segregation of voiced speech
  - Binaural segregation of natural speech
- **Summary**

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)**
- **Our study (Hu & Wang'01) applies different grouping mechanisms in the low-frequency and high-frequency ranges (see Bird & Darwin'97)**
  - Low-frequency signals are grouped based on periodicity and temporal continuity
  - High-frequency signals are grouped based on AM and temporal continuity

AM - Example



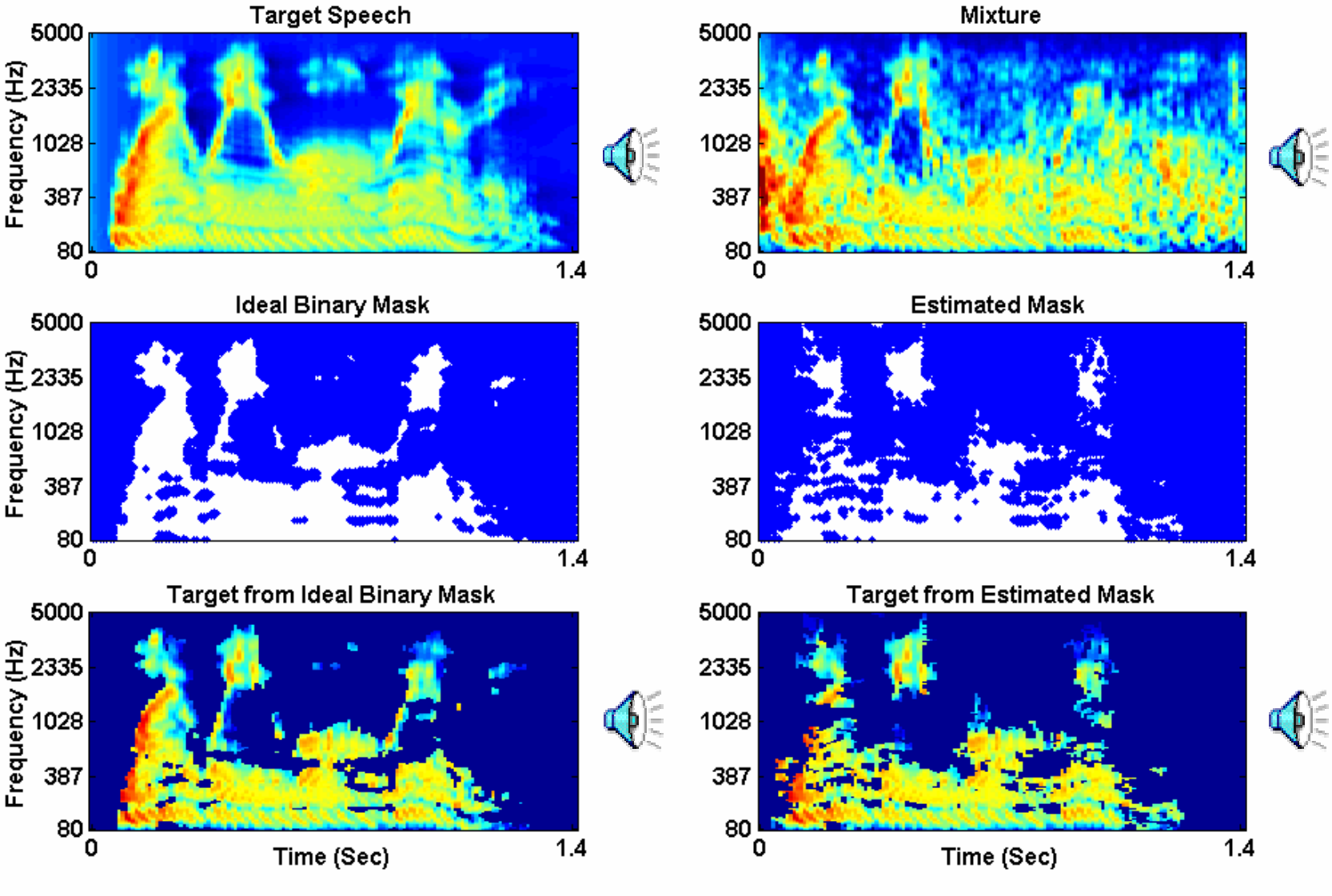
(a) The output of a gammatone filter (center frequency: 2.6 *k*Hz) in response to clean speech

(b) The corresponding autocorrelation function

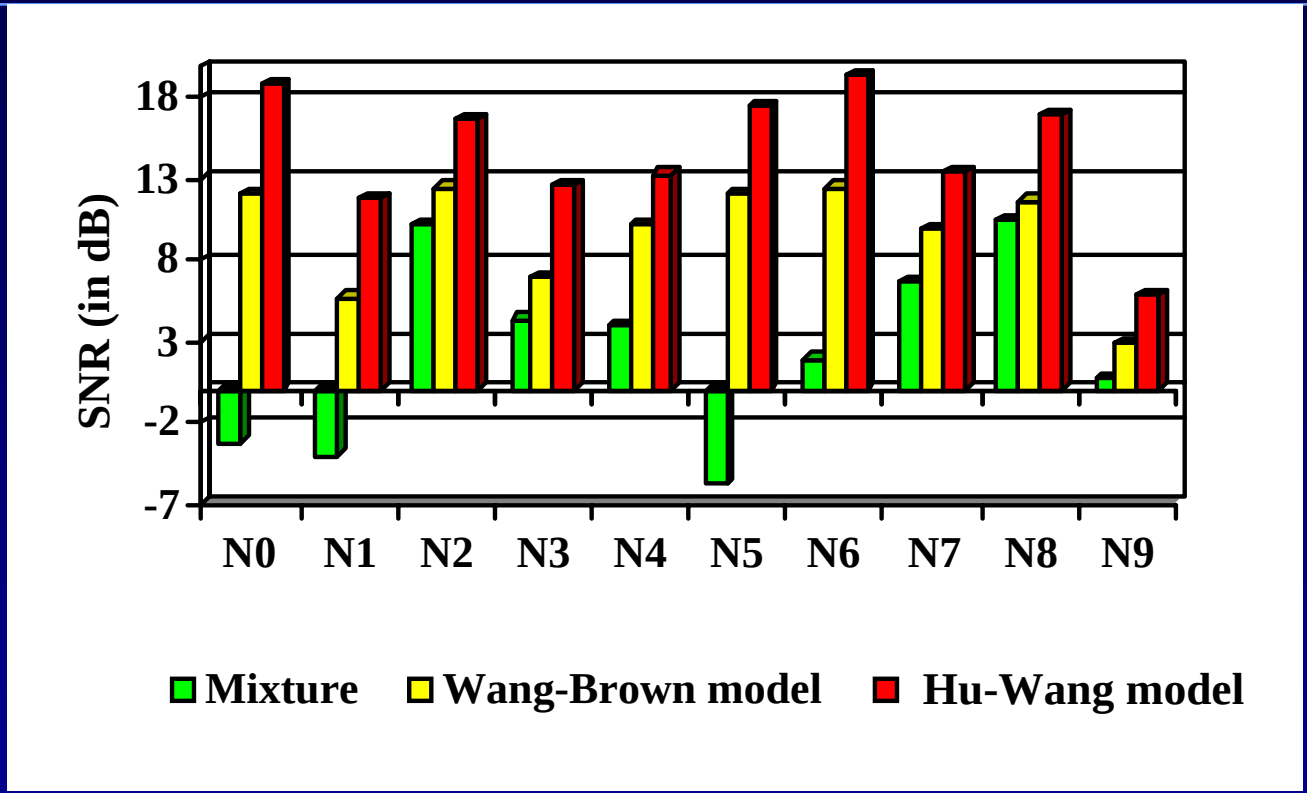
T-F Unit Labeling and Grouping

- **In the low-frequency range, a T-F unit is labeled by comparing its periodicity with the estimated target pitch**
- **In the high-frequency range:**
  - Due to their wide bandwidths, high-frequency filters respond to multiple harmonics. These responses are amplitude modulated due to beats and combinational tones (Helmholtz, 1863)
  - A T-F unit in the high-frequency range is labeled by comparing its AM repetition rate with the estimated target pitch
- **New segments corresponding to unresolved harmonics are formed based on temporal continuity and cross-channel correlation of response envelopes (i.e. common AM). Then they are grouped into the foreground stream according to AM repetition rates**

Voiced Speech Segregation Example



Systematic SNR Results

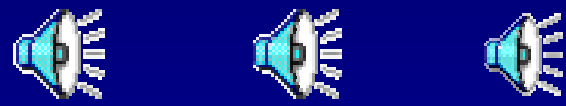


- Evaluation on a corpus of 100 mixtures (Cooke'93): 10 voiced utterances x 10 noise intrusions based on ideal binary masks
- Average SNR gain: 12.1 dB; 5 dB better than the Wang-Brown model (1999)

Segregation Examples

Mixture 

Ideal Binary Mask 

Estimated Binary Mask 

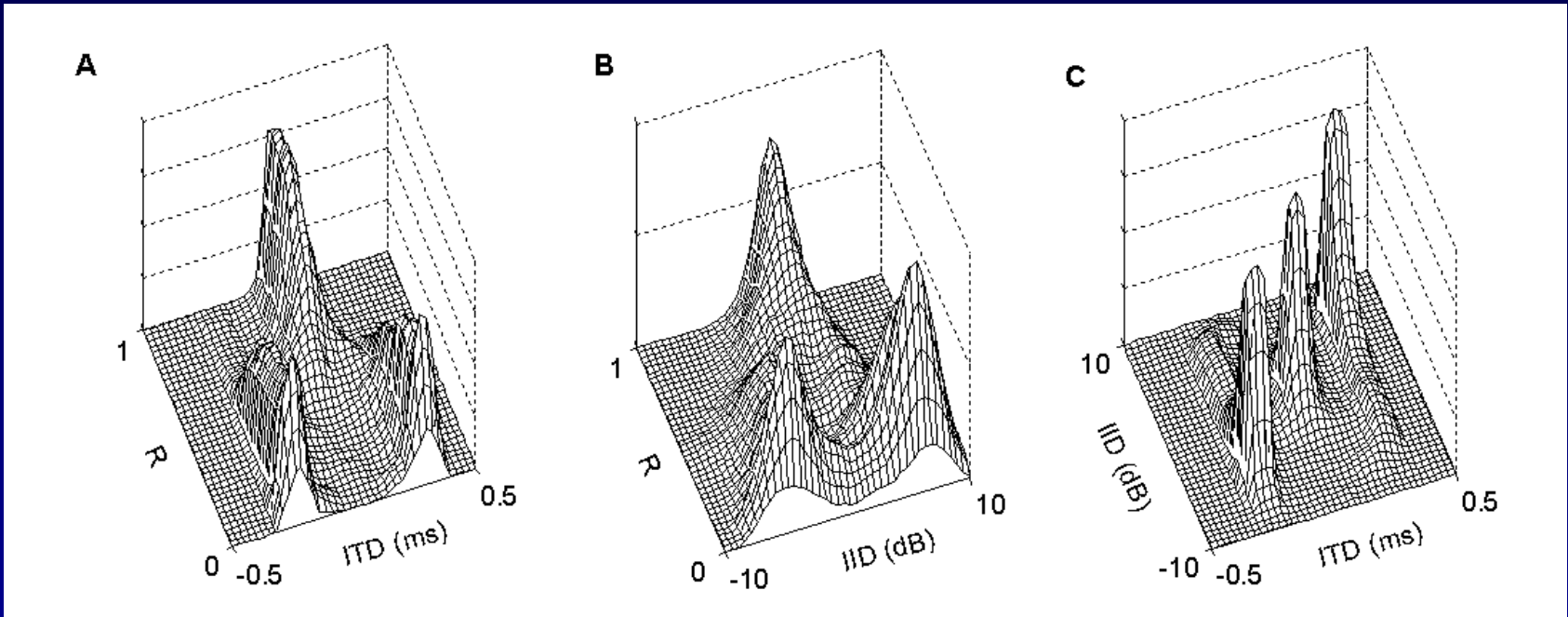
Binaural Segregation of Natural Speech

- **The objective is to model a listener's ability to selectively attend to one talker while filtering out acoustic interference using binaural cues**
- **Binaural speech segregation is applicable to both voiced and unvoiced speech**
- **Our study (Roman, Wang, & Brown'03) focuses on location cues:**
  - Interaural time difference (ITD)
  - Interaural intensity difference (IID)
- **Again, the computational goal is to estimate ideal binary masks**

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 original signals changes. This interaction produces characteristic clustering in the joint ITD-IID space**
- **The core of our model lies in deriving the statistical relationship between the relative strength and the binaural cues**
  - Independent supervised learning for different spatial configurations and different frequency bands in the joint ITD-IID space
- **The model yields large SNR improvements**
  - For 2-source configurations, average SNR gain (at the better ear) ranges from 13.7 dB to 5 dB depending on azimuth separation and deviation from median plane
  - For 3 sources, average SNR gain is 11.3 dB in good configurations

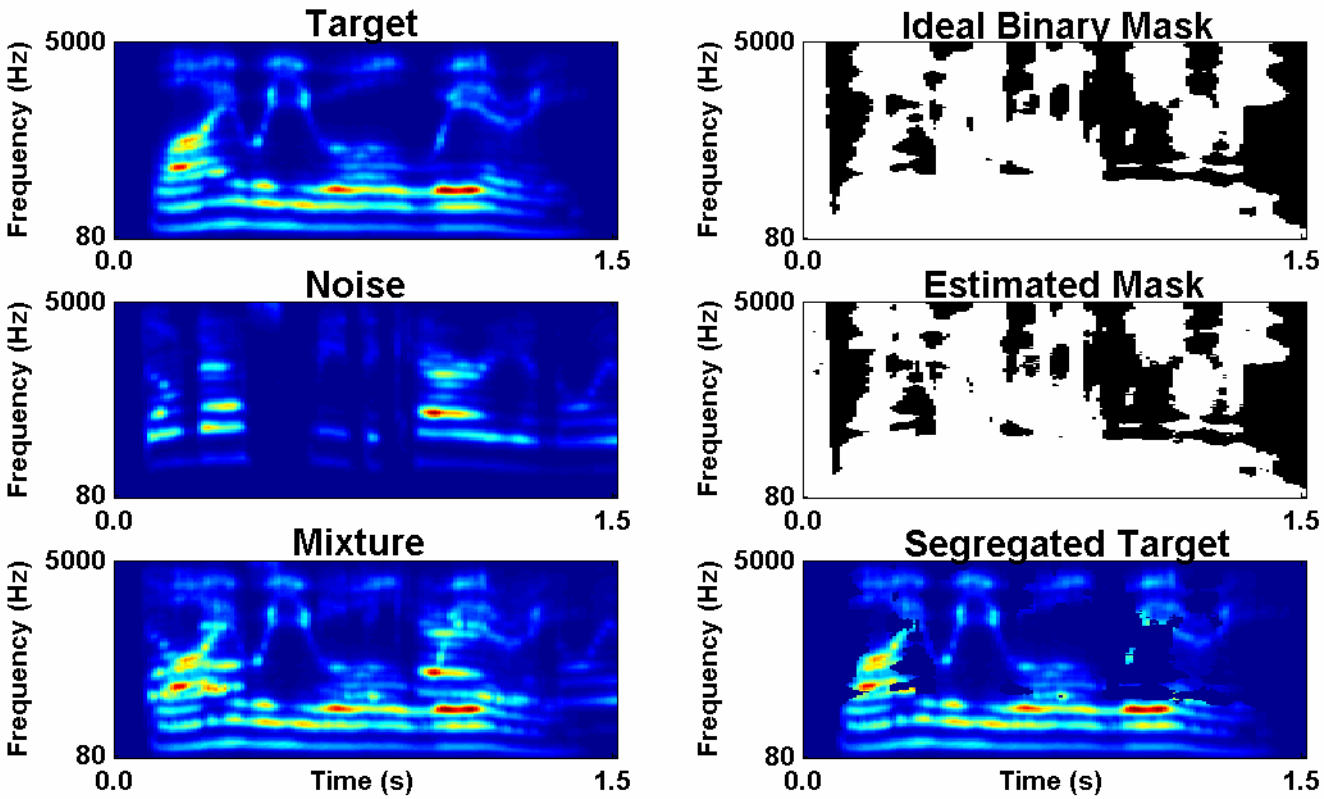
3-Source Configuration Example



- Data histograms for one channel (center frequency: 1.5 kHz) from speech sources with target at 0^{o} and two intrusions at -30^{o} and 30^{o} (*R*: relative strength)

- Clustering in the joint ITD-IID space

Example (Target: 0^{o}, Noise: 30^{o})



Target Noise Mixture Ideal binary mask Result

Sound Demos

2 sound sources (Target: 0°, Noise: 30°)

Target

Noise	Mixture	Segregated target
'Cocktail Party'		
Siren		
Female Speech		

3 sound sources (Target: 0°, Noise1: -30°, Noise2: 30°)

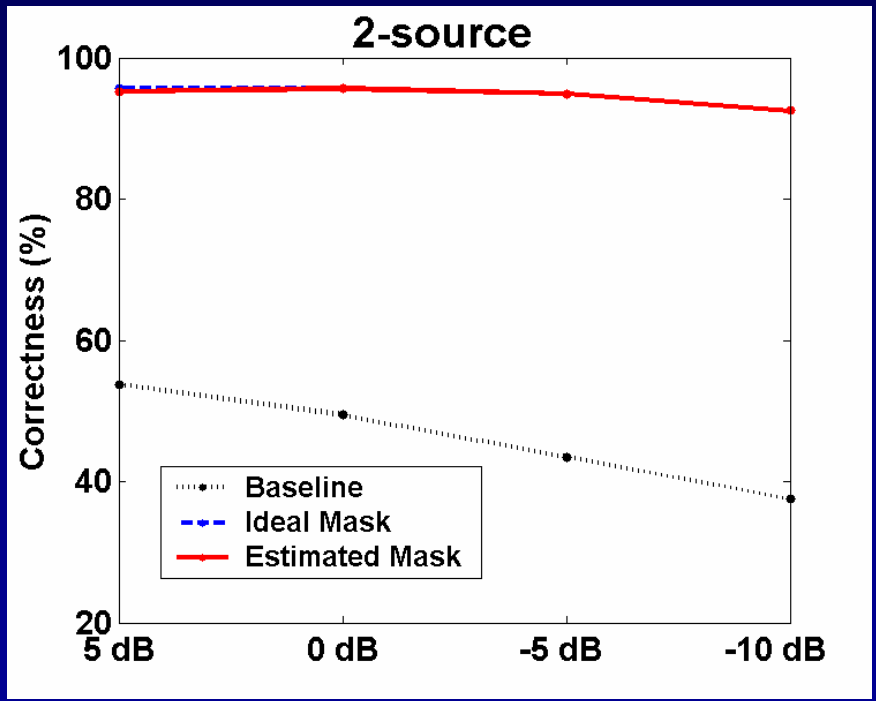
Target

Noise2

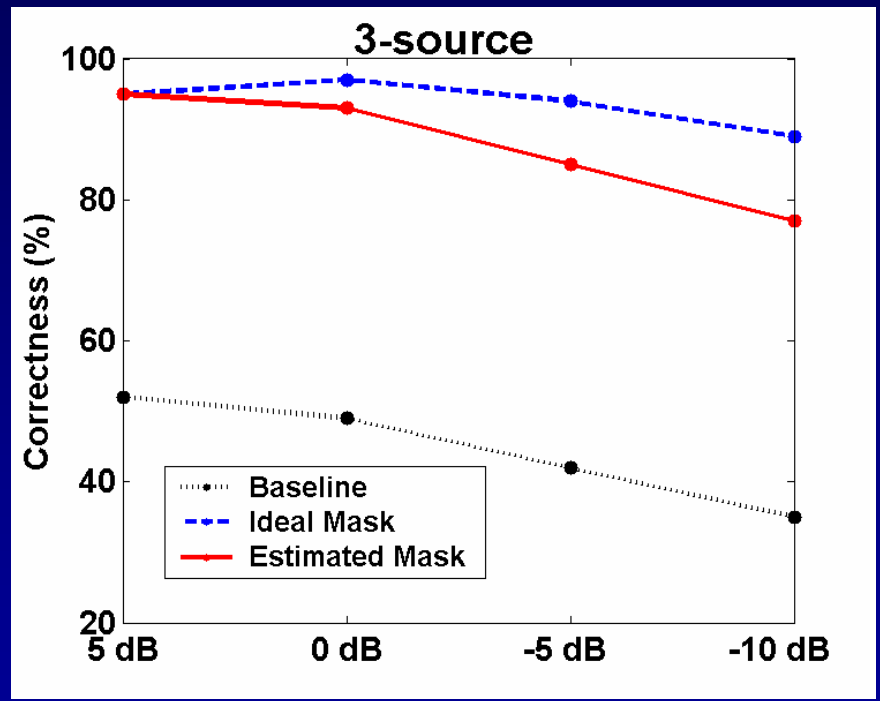
Noise1	Mixture	Segregated target
'Cocktail Party'		
Female Speech		

ASR Evaluation

- **We employ the missing-data technique for robust speech recognition (Cooke *et al.*'01). The task domain is recognition of connected digits**



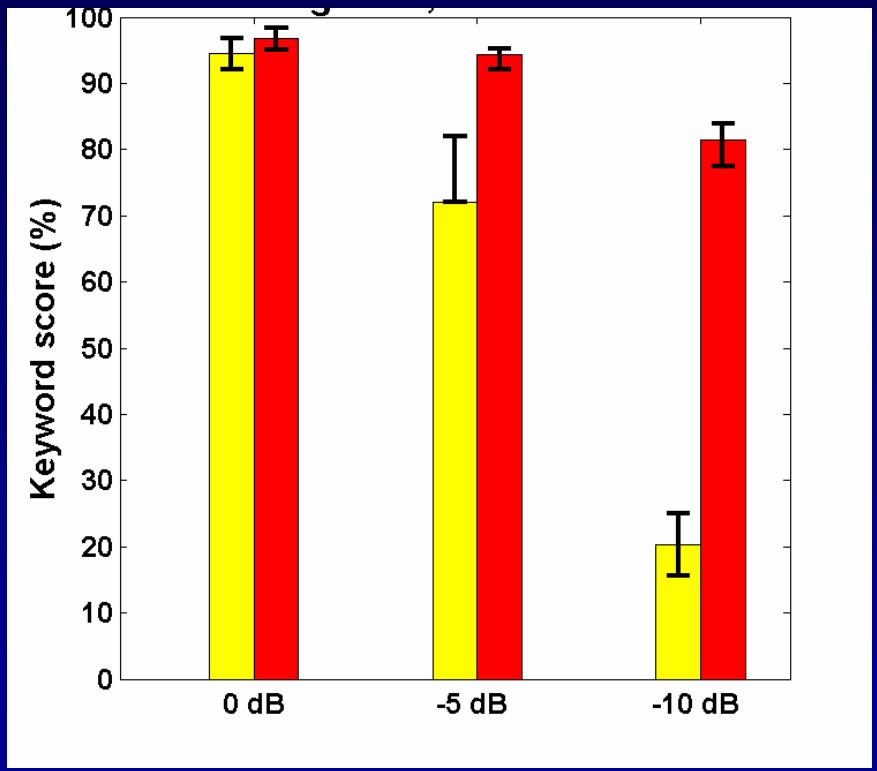
Target at 0^{o}
Intrusion (male speech) at 30^{o}

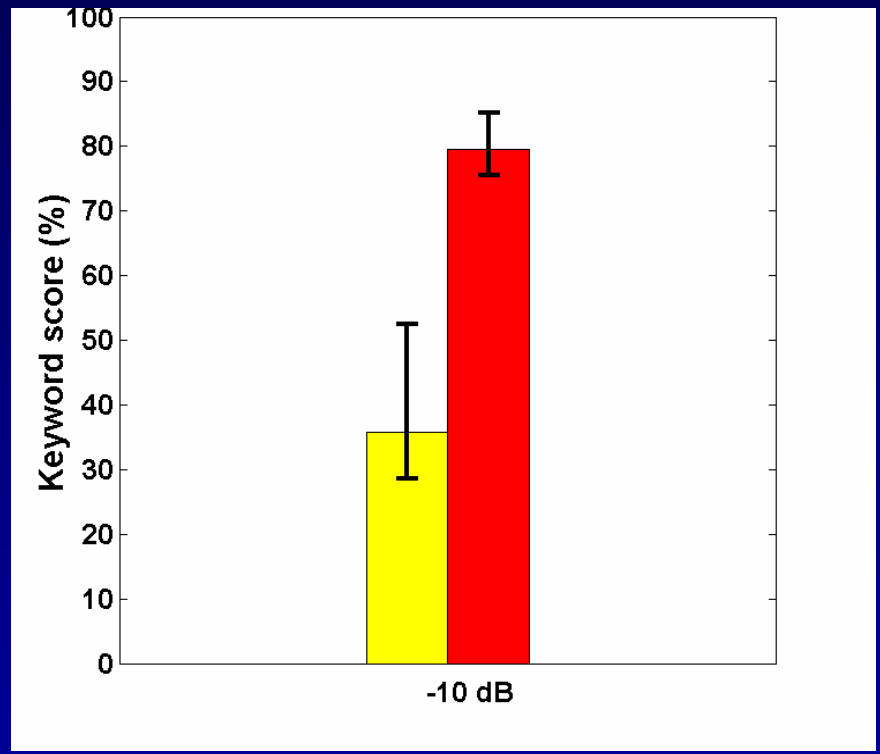


Target at 0^{o}
Two intrusions at 30^{o} and -30^{o}

Speech Intelligibility Evaluation

- **We employ the Bamford-Kowal-Bench sentence database that contains short semantically predictable sentences as target**





Mixture **Segregated**

Two-source (0^o, 5^o) condition
Interference: babble noise

Three-source (0^o, 30^o , -30^o) condition
Interference: male utterance & female utterance

Summary

- **A clear understanding of the computational goal of ASA is important for model development**
  - Computational theory analysis
  - Evaluation criteria for CASA
- **Discussion of different CASA objectives**
- **Ideal binary mask as a putative goal**
  - Example studies estimate ideal binary masks for monaural and binaural speech segregation

Acknowledgement

- **Research supported by AFOSR and NSF**