

USING COMPUTATIONAL MODELS OF BINAURAL HEARING TO IMPROVE AUTOMATIC SPEECH RECOGNITION: Promise, Progress, and Problems

Richard Stern

**Department of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213**

**Telephone: (412) 268-2535; FAX: (412) 268-3890
Email: rms@cs.cmu.edu; <http://www.ece.cmu.edu/~rms>**

**AFOSR Workshop on Computational Audition
August 9, 2002**

Introduction - using binaural processing to improve speech recognition accuracy

- **We're doing better than I expected** but there is still a lot to be done
- **Talk outline**
 - Briefly review some relevant binaural phenomena
 - Talk briefly about “classical” modeling approaches
 - Discuss some specific implementations of binaural processors that are particularly relevant to automatic speech recognition (ASR)
 - Comment a bit on results and future prospects

I'll focus on ...

■ Studies that

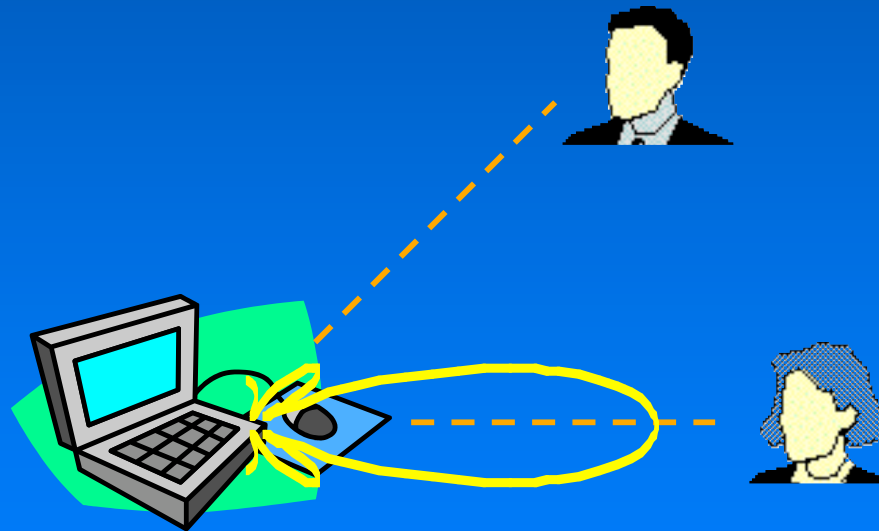
- are based on the binaural temporal representation and are
- applied in a meaningful fashion to automatic speech recognition

■ Will not consider ...

- Other types of multiple microphone processing
 - » Fixed and adaptive beamforming
 - » Noise cancellation using adaptive arrays
- Applications to source localization and separation
- Applications to hearing impaired
- Other ASR application using other approaches (like sub-band coherence-based enhancement)

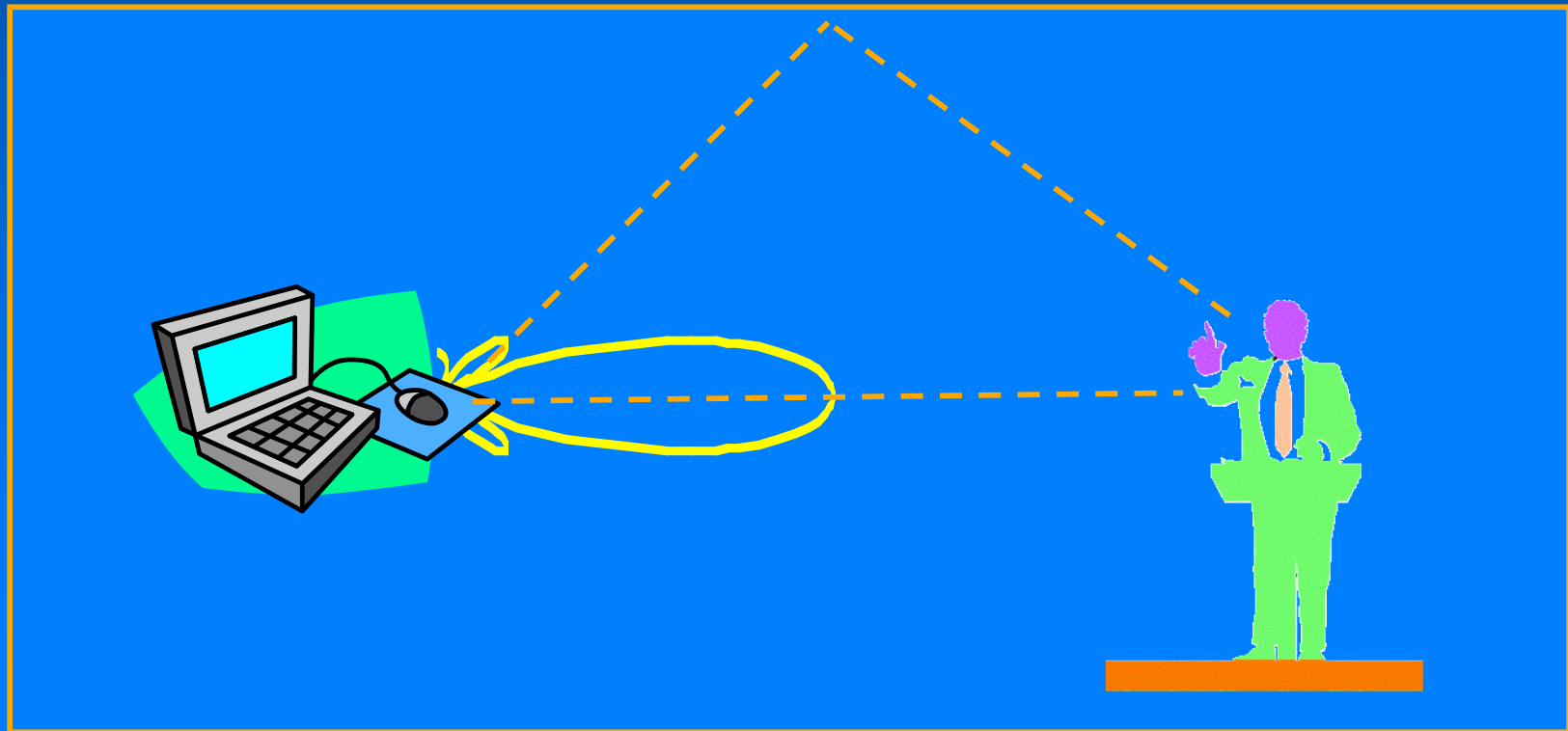
How can two (or more) ears help us?

- The binaural system can **focus attention** on a single speaker in a complex acoustic scene



Another big binaural advantage ...

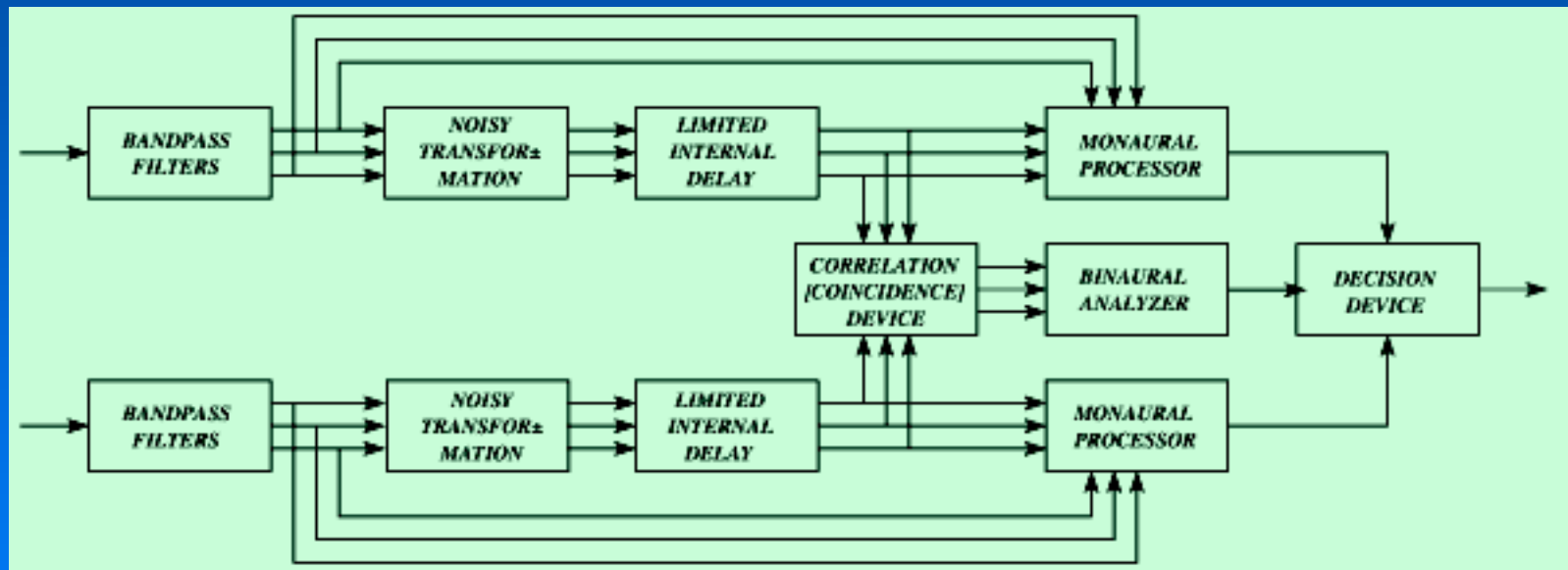
- The binaural system can **suppress reflected components** of sound in a reverberant environment



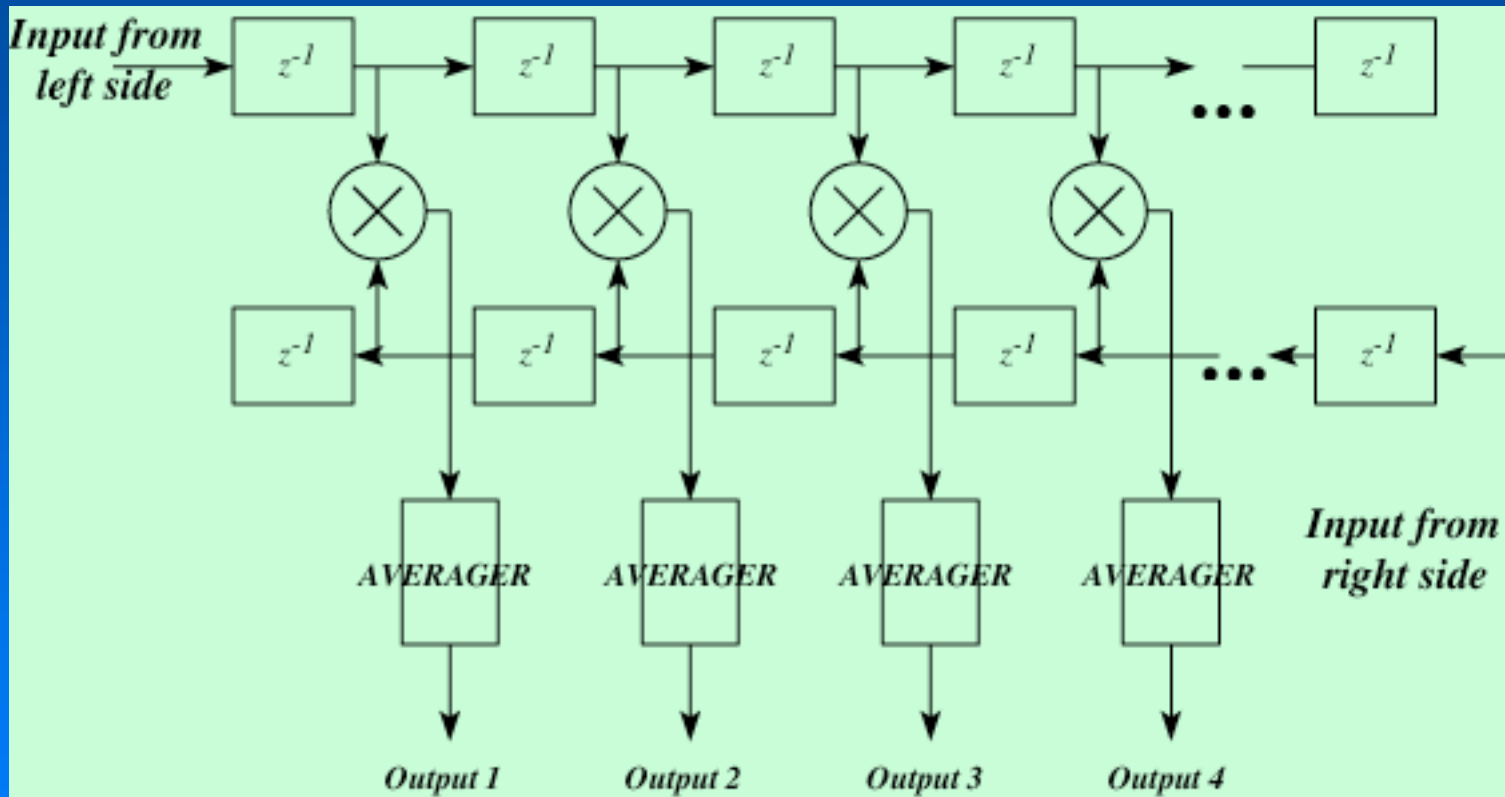
Primary binaural phenomena

- Interaural time delays (ITDs)
- Interaural intensity differences (IIDs)

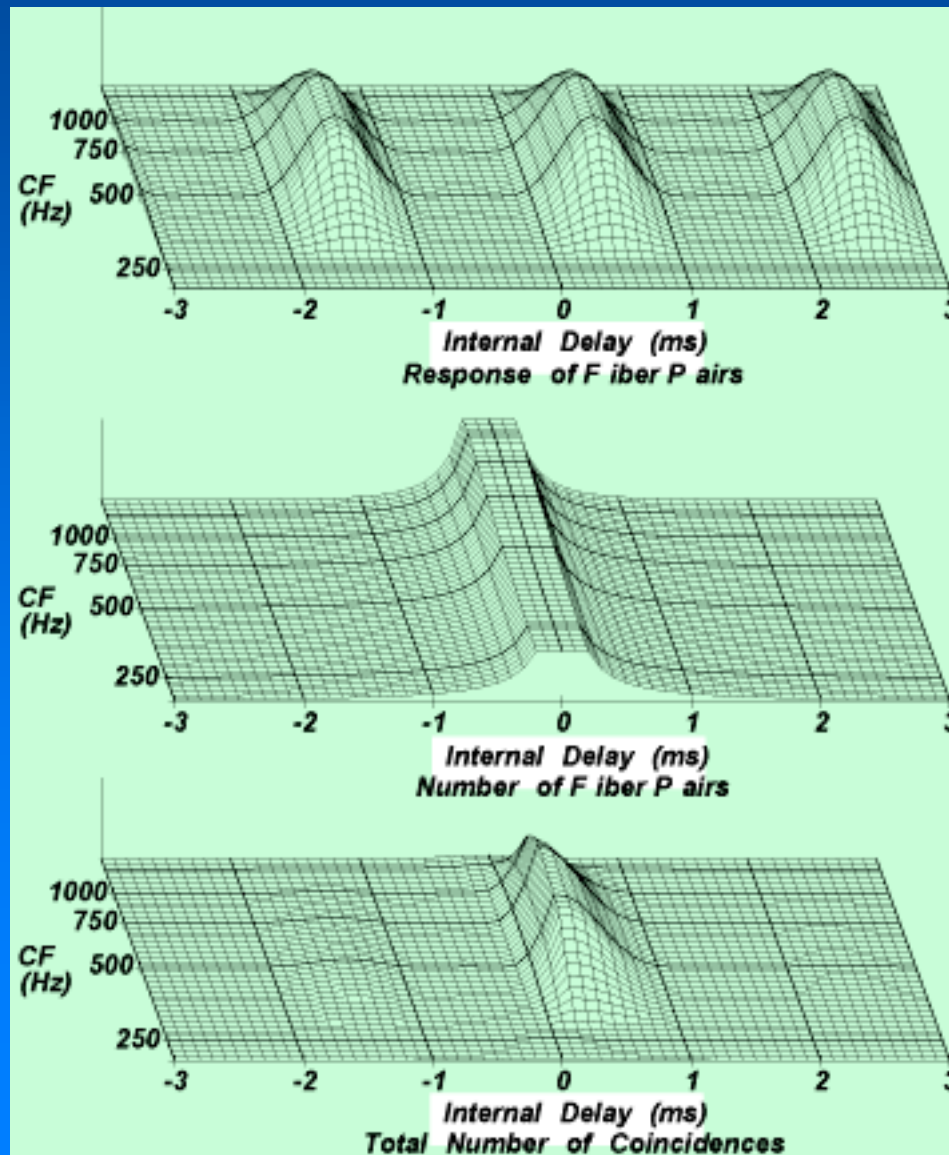
The classical model of binaural processing (Colburn and Durlach, 1978)



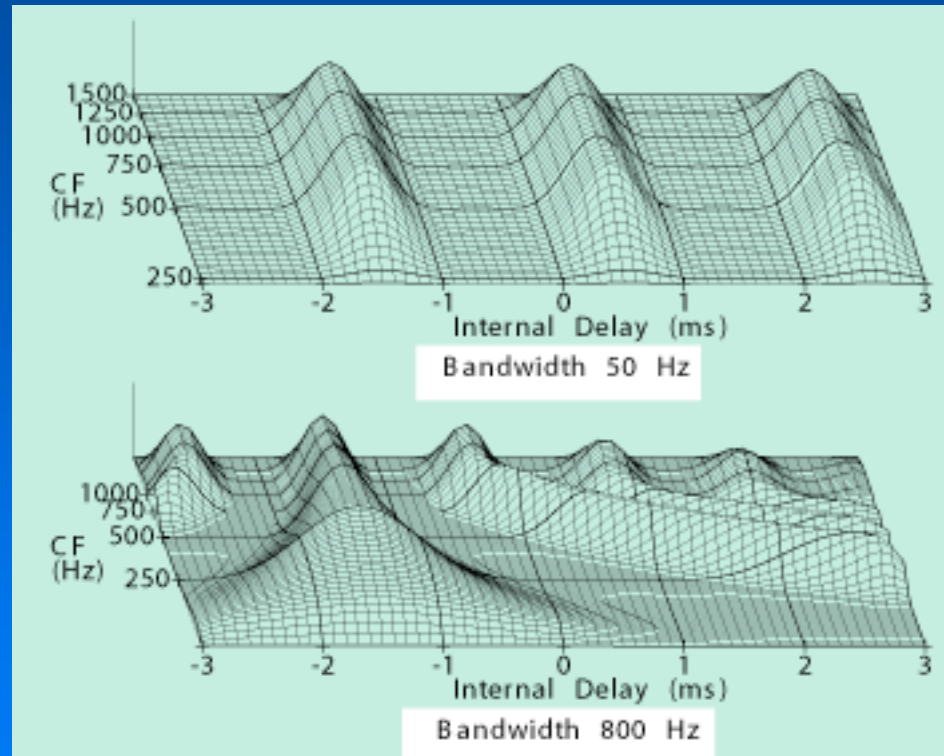
Jeffress's model of ITD extraction (1948)



Response to a 500-Hz tone with $-1500\text{-}\mu\text{s}$ ITD



Response to 500-Hz noise with $-1500\text{-}\mu\text{s}$ ITD



Other important binaural phenomena

■ Binaural “sluggishness”

- Temporal integration “blurs” effects of running time

■ Information from head-related transfer functions (HRTFs)

- Impetus for significant work in externalization and virtual environments
- Also enables analysis of relationships between ITDs and IIDs

■ The precedence effect

- First-arriving wavefront has greatest impact on localization

■ Secondary “slipped-cycle” effects and “straightness weighting”

- Localization mechanisms also responsive to consistency over frequency

Some of the groups involved

- **Pittsburgh** - Stern, Sullivan, Palm, Raj, Seltzer et al.
- **Bochum** - Blauert, Lindemann, Gaik, Bodden et al.
- **Oldenburg** - Kollmeier, Peissig, Kleinschmidt, Schortz et al.
- **Dayton** - Anderson, DeSimio, Francis
- **Sheffield** - Green, Cooke, Brown, (and Ellis, Wang, Roman) et al.

Typical elements of binaural models for ASR

■ Peripheral processing

- HRTFs or explicit extraction of ITDs and IIDs vs. frequency band

■ Model of auditory transduction

- Prosaic (BPF, rectification, nonlinear compression) or AIM

■ Interaural timing comparison

- Direct (cross-correlation, stereoausis, etc.) or enhanced for precedence (a la Lindemann)

■ Time-intensity interaction

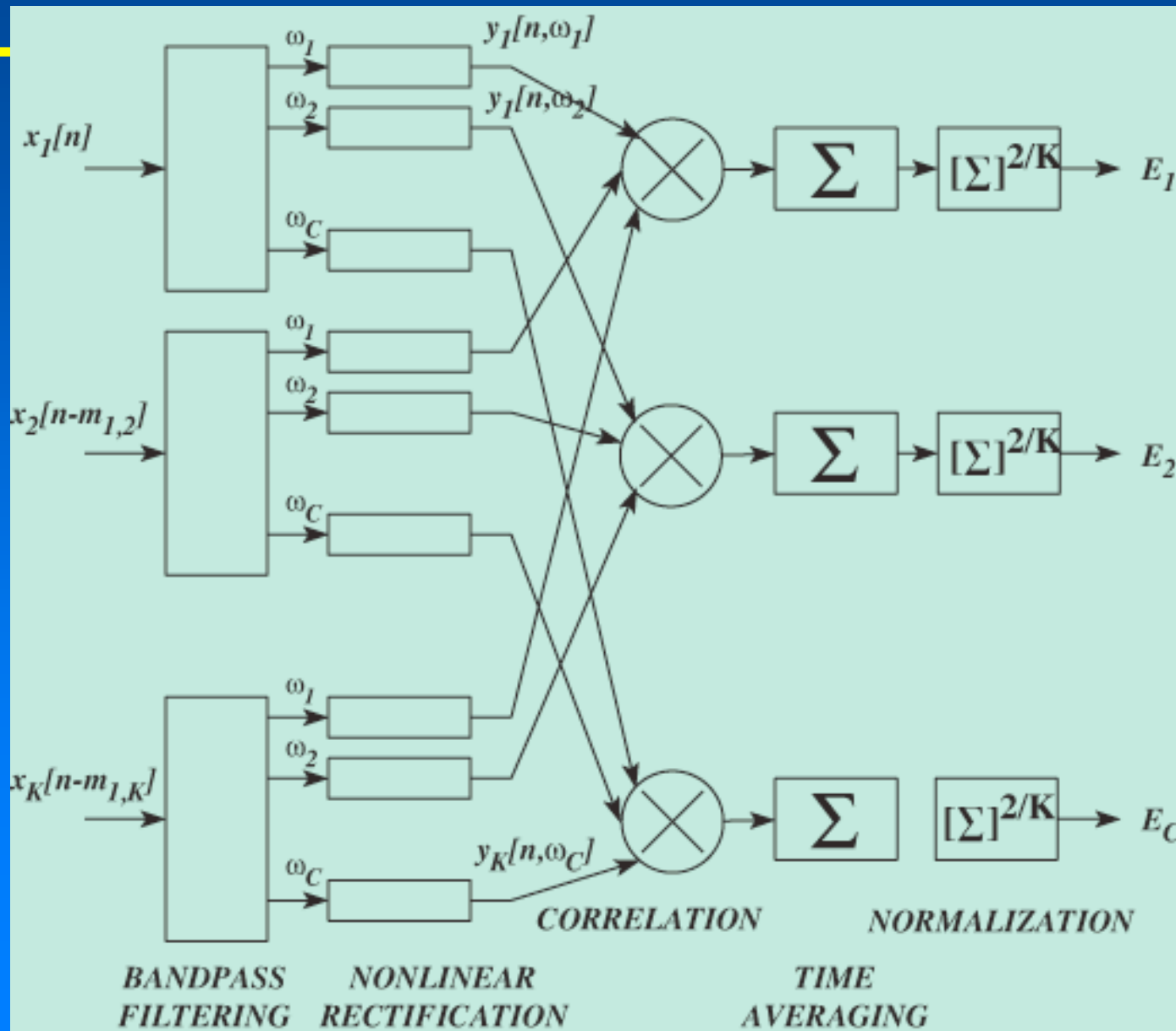
- Use of interaural intensity information to reinforce/vitiate temporal information (e.g. Gaik, Peissig)

■ Possible restoration of “missing” features

■ Feature extraction of enhanced display

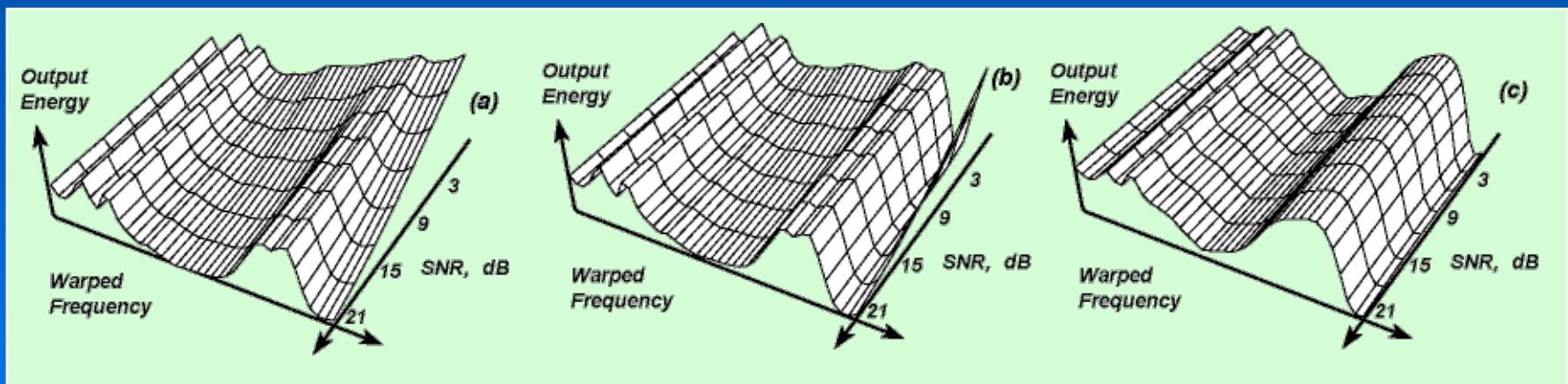
■ Decision making (Bayesian or using neural networks)

Some (old) work from CMU: correlation-based ASR motivated by binaural hearing



The good news: vowel representations improved by correlation processing

■ Reconstructed features of vowel /a/



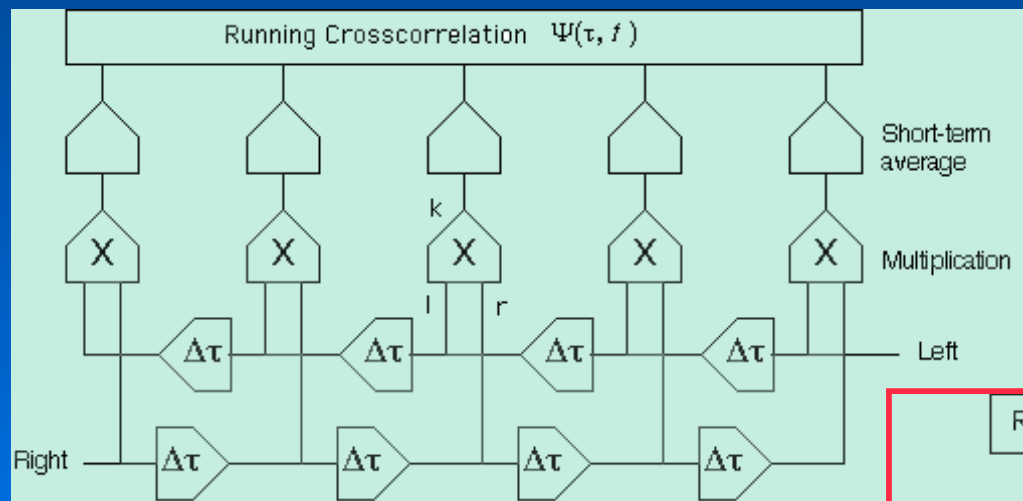
Two inputs
zero delay

Two inputs
120- μ s delay

Eight inputs
120- μ s delay

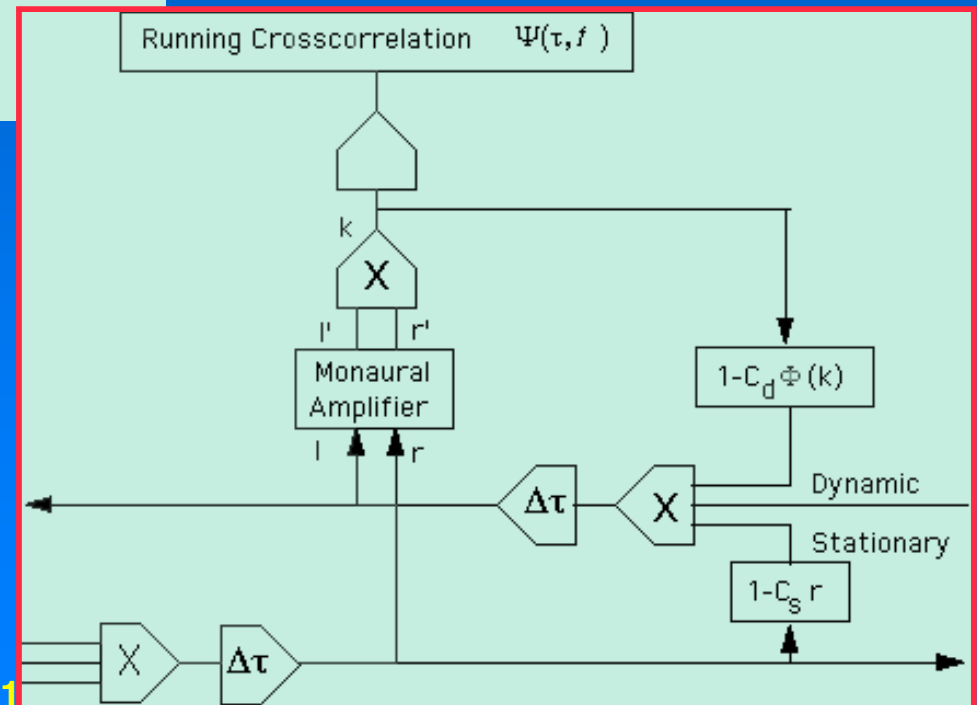
- But the bad news is that error rates in real environments go down only a small amount, with a lot more processing

The Lindemann model to accomplish the precedence effect



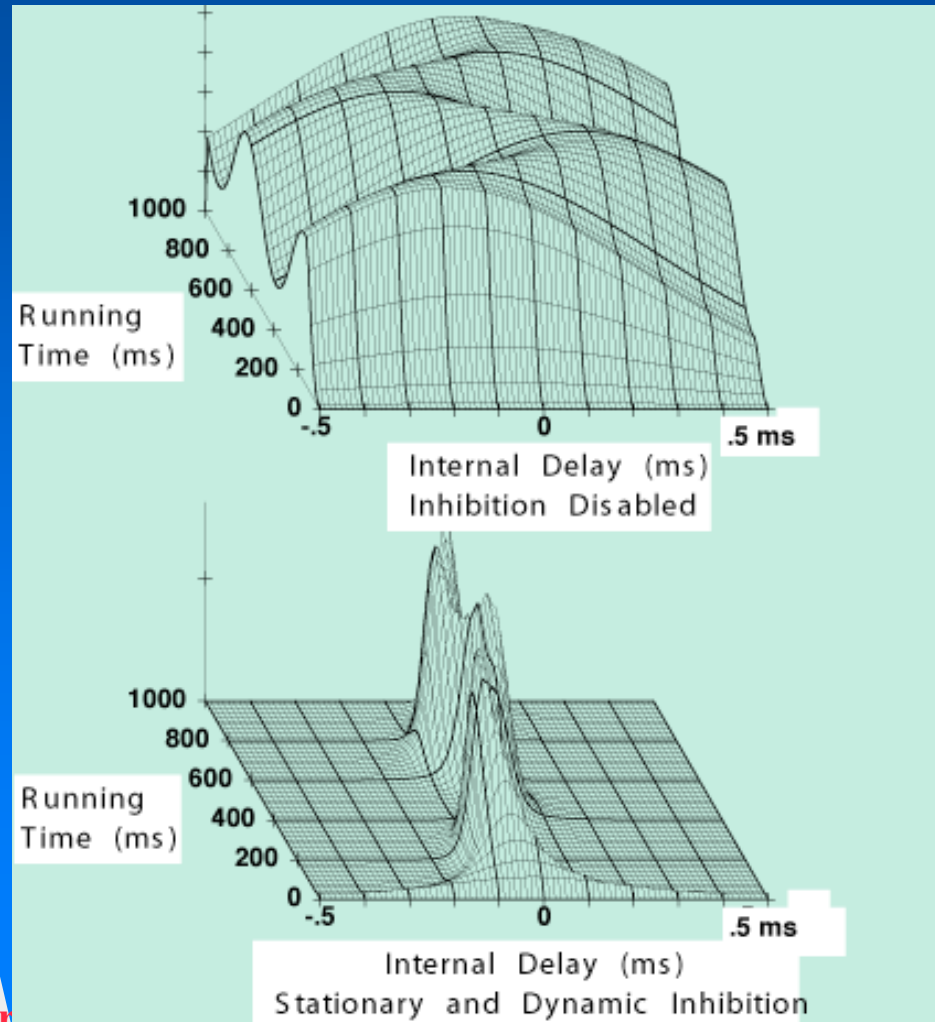
Lindemann inhibition

Blauert cross-correlation



Slide 1

Sharpening effect of Lindemann inhibition



Comment: Also observe precedence phenomena (as expected) and a natural time-intensity trade.

Other techniques use by the Bochum group

■ Gaik

- Collected statistics of ITDs and IIDs of signals through HFTF filters
- Used statistics to estimate joint pdf of ITD and IID, conditioned on source location

■ Bodden

- Detected source location and implemented source separation algorithm by differentially weighting different frequency bands

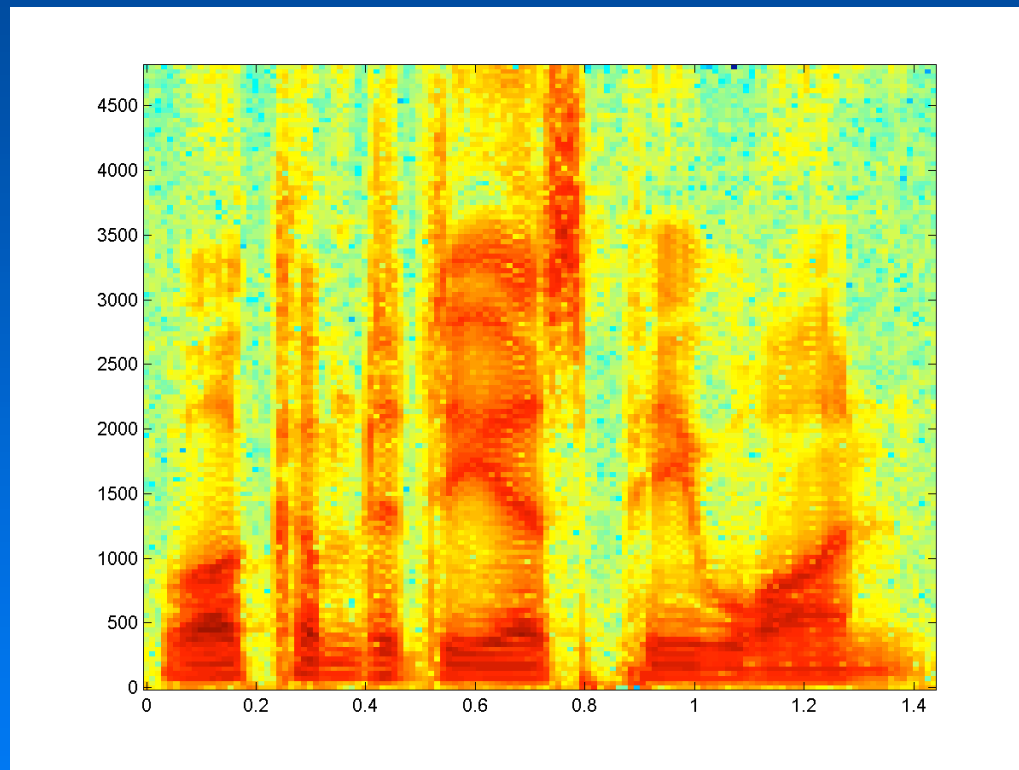
- **Comment:** Oldenburg group has developed a similar model (that differs in many details), but without the Lindemann inhibition

Missing-feature recognition

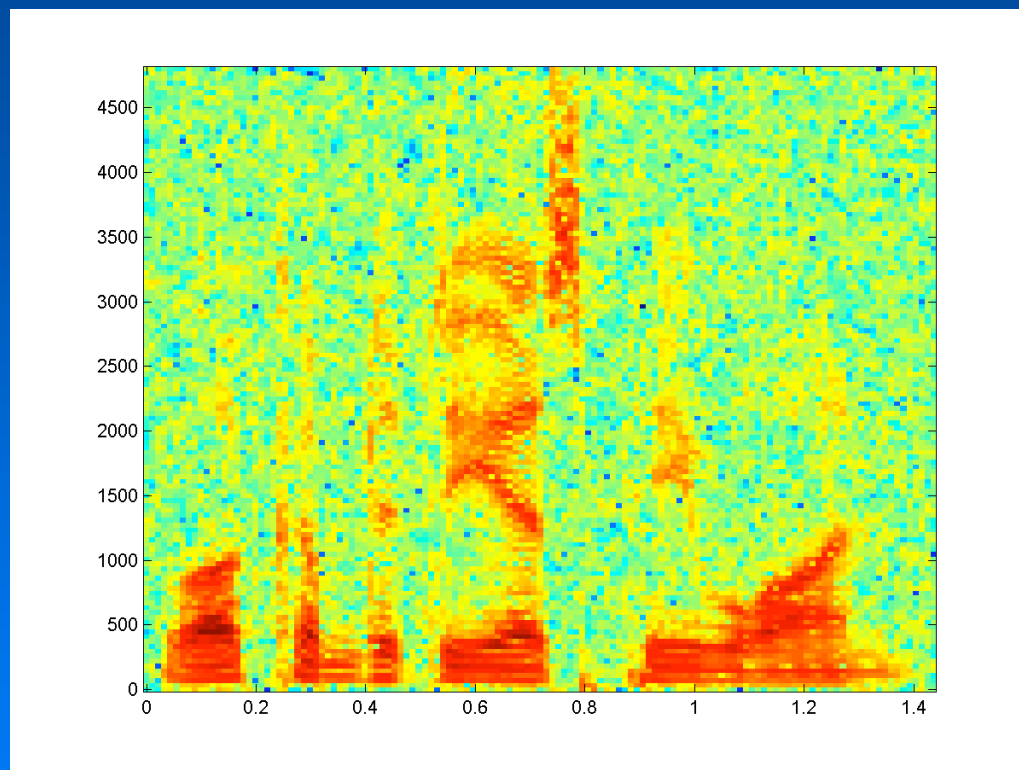
■ General approach:

- Determine which cells of a spectrogram-like display are unreliable (or “missing”)
- Ignore missing features or make best guess about their values based on data that are present

Original speech spectrogram

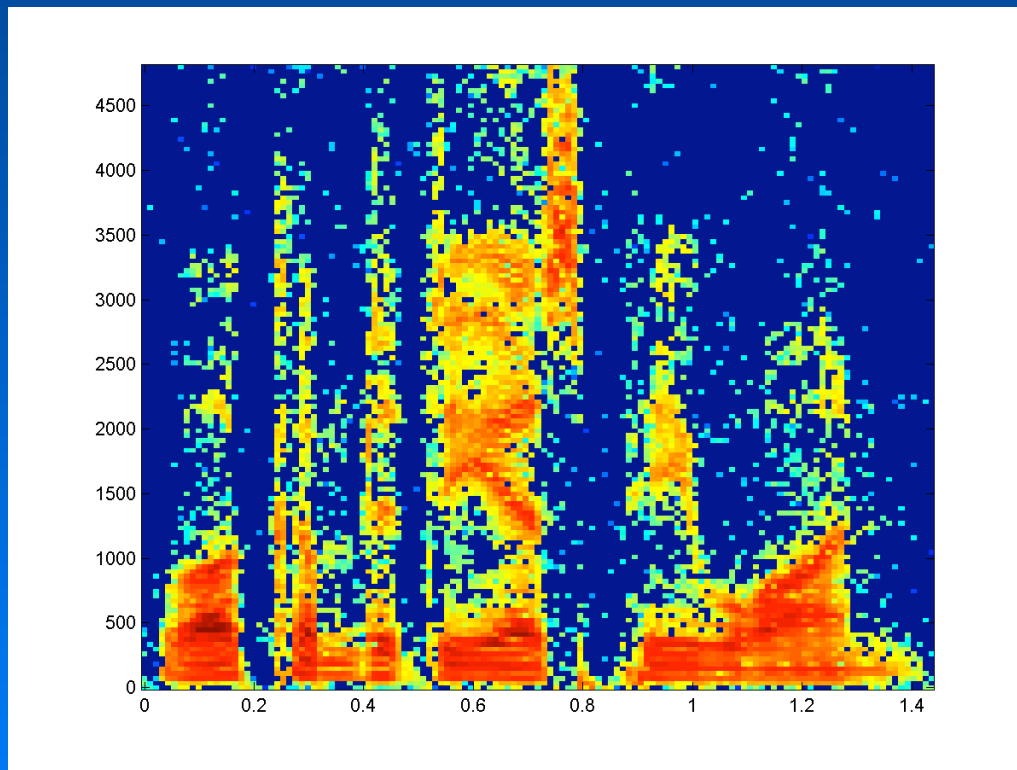


Spectrogram corrupted by white noise at SNR 15 dB



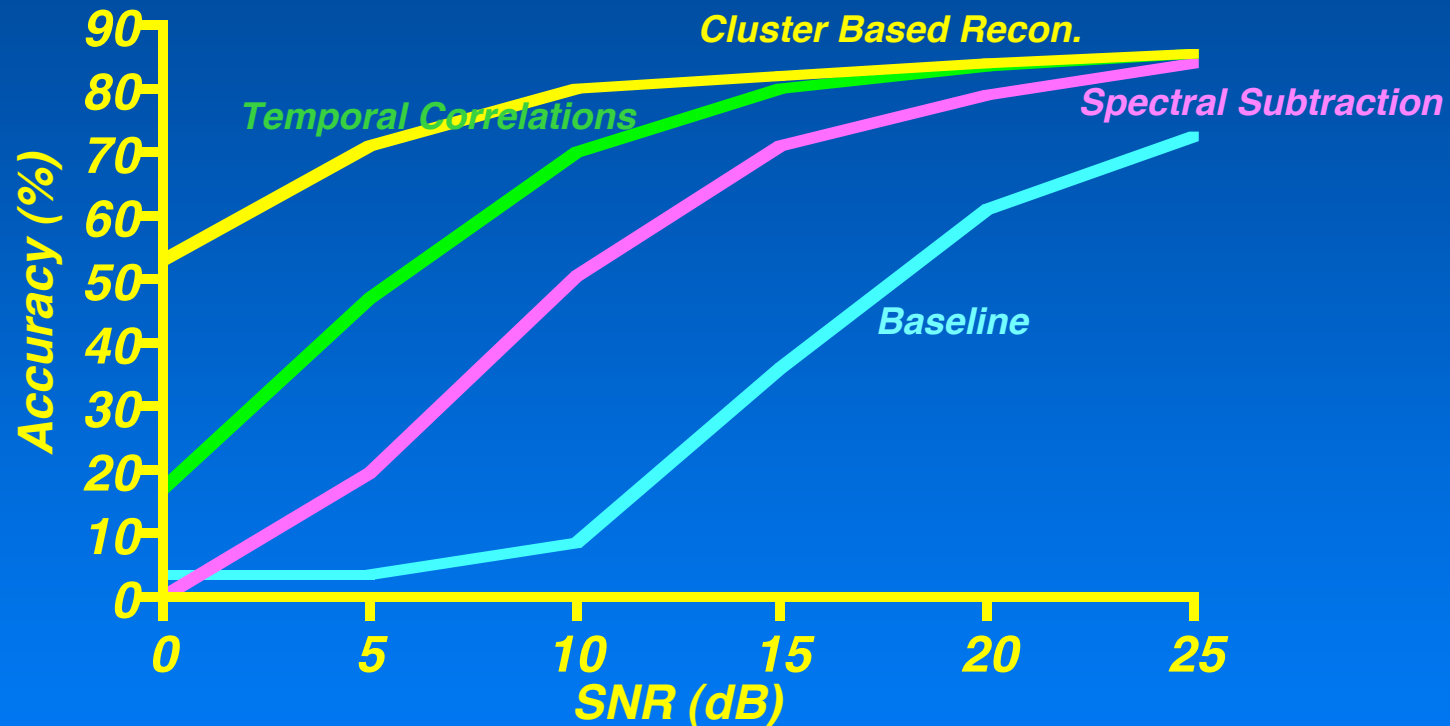
- Some regions are affected far more than others

Ignoring regions in the spectrogram that are corrupted by noise



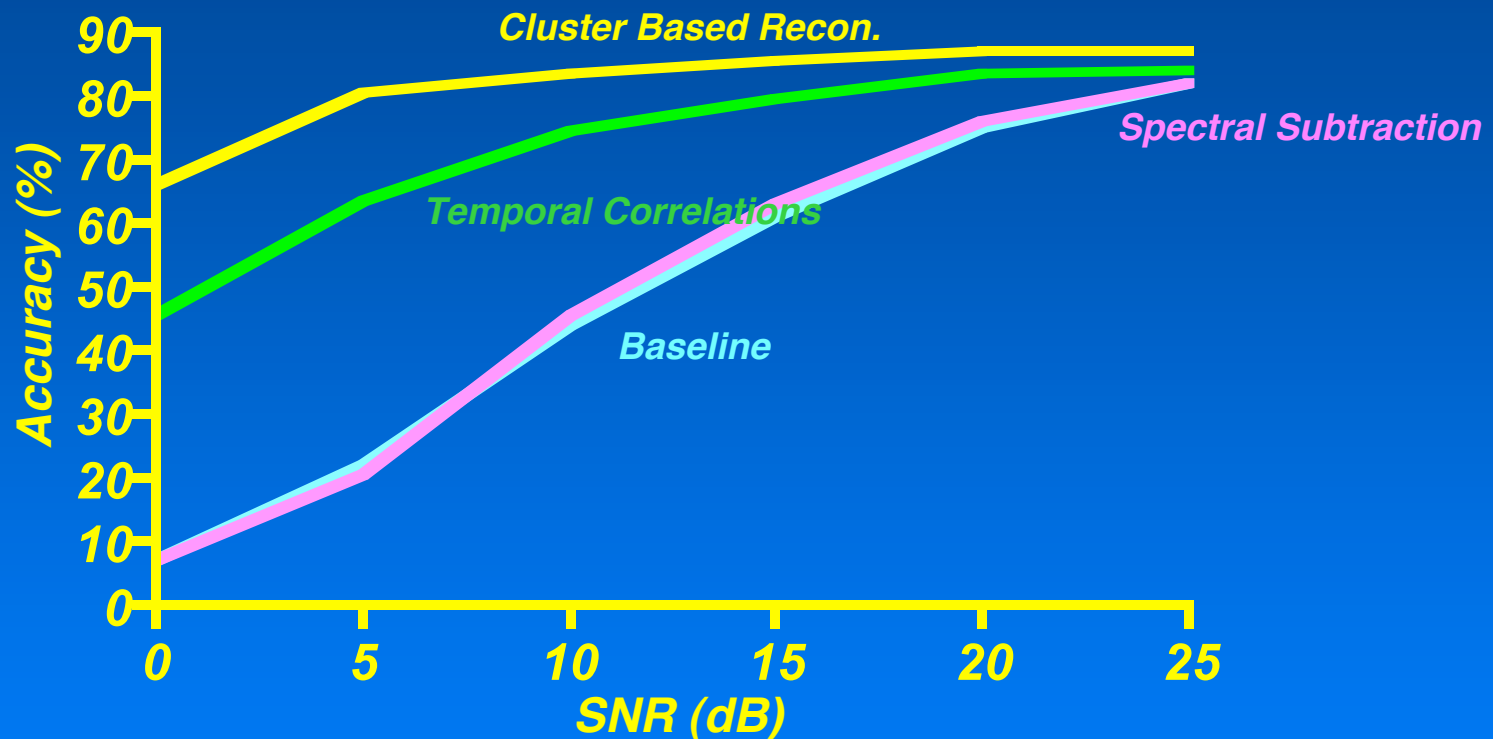
- All regions with SNR less than 0 dB deemed missing (dark blue)
- Recognition performed based on colored regions alone

Recognition accuracy using compensated cepstra, speech corrupted by white noise



- Large improvements in recognition accuracy can be obtained by reconstruction of corrupted regions of noisy speech spectrograms
- Knowledge of locations of “missing” features needed

Recognition accuracy using compensated cepstra, speech corrupted by music



- Recognition accuracy goes up from 7% to 69% at 0 dB with cluster based reconstruction

Latest system from the Oldenburg group

■ Peripheral processing:

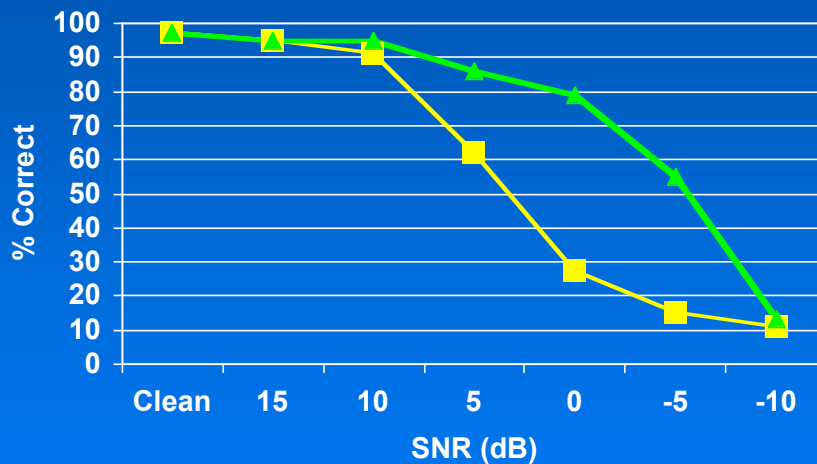
- Gammatone filters
- Envelope extraction, lowpass filtering
- Nonlinear temporal adaptation
- Lowpass filtering

■ Binaural processing:

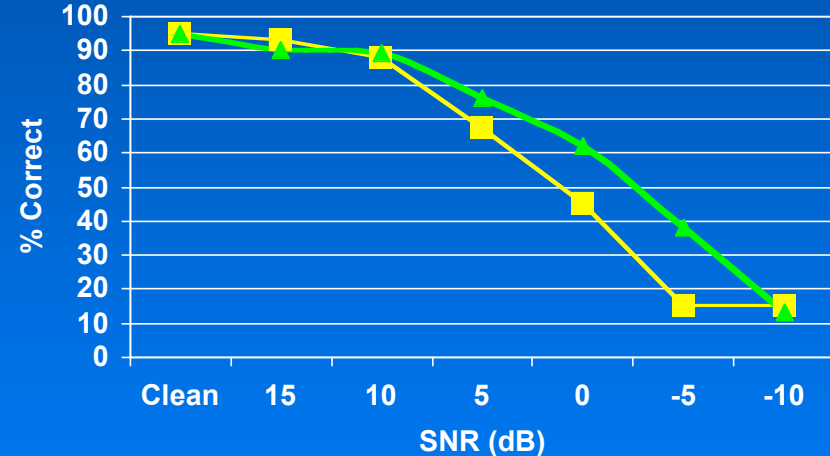
- Direct running cross-correlation (no inhibition)
- Learning of corresponding ITD, IID using a neural network
- Feature extraction from representation in “look direction”

Sample results from the Oldenburg group (Kleinschmidt *et al.* 2001)

■ Anechoic environment:



■ “Moderate” reverberation:



■ **Comment:** System performs worse in reverberation

Some systems developed by the Dayton group

■ Binaural Auditory Image Model (BAIM):

- HRTFs
- Auditory image model (AIM)
- Cross-correlation with and without Lindemann inhibition
- ITD/IID comparison using Kohonon self-organizing feature map

■ Cocktail-party Processor (1995):

- HRTFs
- Conventional peripheral processing with Kates model
- Cross-correlation with Lindemann inhibition

[BAIM worked somewhat better for most conditions]

Some comments, kudos, and concerns ...

- **Be very skeptical with results obtained using artificially added signals and noise! Nevertheless some progress has definitely been made.**
 - Digitally adding noise almost invariably inflates performance
 - Use of room image models to simulate reverberant room acoustics may be more reasonable
- **Lots of information is being ignored in many current models**
 - Synchrony info a la Seneff, Ghitza;
 - Complex timbre information as suggested by Lyon, Slaney?
- **The Lindemann model may not be the best way to capture precedence**
- **Missing feature approaches should be very promising**
- **Too much computational modeling and not enough insight into fundamental processes**

Summary

- Binaural processing has the ability (in principle) to improve speech recognition accuracy by providing spatial filtering and by combating the effects of room reverberation.
- Current systems are realizing some gains but are just now beginning to realize that promise.
- Faster and more efficient computation will be a real spur for research in this area over the next five years.

