SIGNAL PROCESSING FOR SOUND SEPARATION AND ROBUST REPRESENTATION

Richard Stern

Robust Speech Recognition Group
Carnegie Mellon University

Telephone: (412) 268-2535
Fax: (412) 268-3890
rms@cs.cmu.edu
http://www.cs.cmu.edu/~rms

AFOSR/NSF Symposium on Speech Separation
November 4, 2004

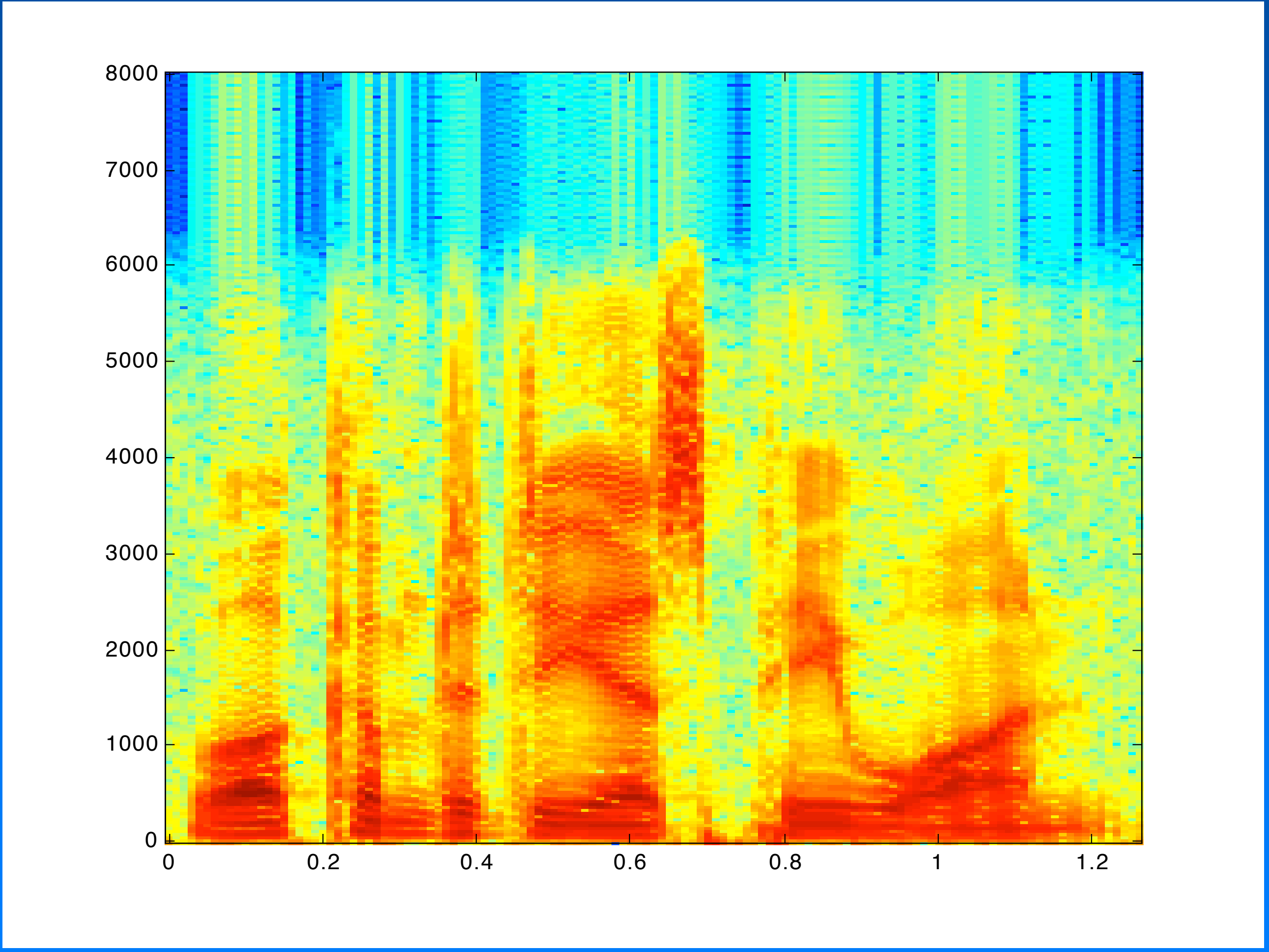
Outline of talk

- Goals: Review and discuss some major issues in signal representation for robust recognition and signal separation
- Current issues in basic peripheral auditory representation
- Classical problems in robust speech recognition
- Generations of solutions to representation and separation problems
  - "Classical" solutions
  - "Transitional" solutions
  - Solutions based on auditory scene analysis

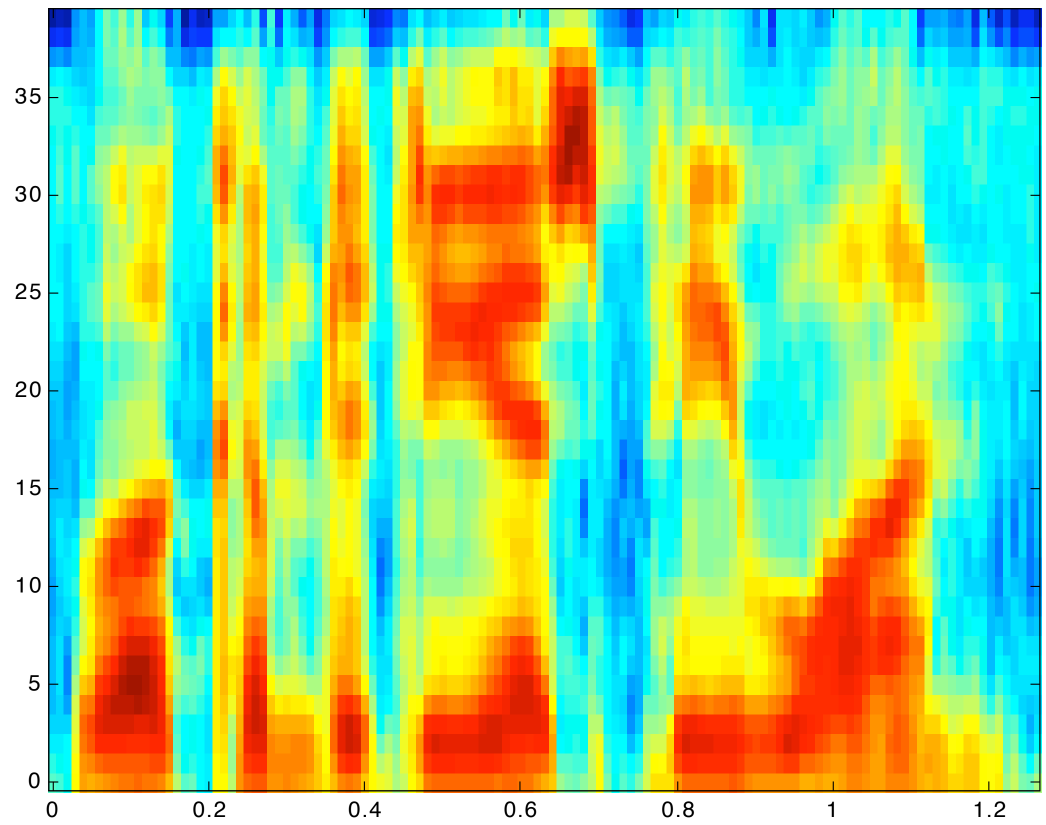
What speech recognizers most frequently see: Mel frequency cepstal coefficients (MFCCs)

- Apply Hamming windows to segment waveform into frames
- Compute frequency response for each frame using DFTs
- Multiply magnitude of frequency response by triangular weighting functions to produce 25-40 channels
- Compute log of weighted magnitudes for each channel
- Take inverse discrete cosine transform (DCT) of weighted magnitudes for each channel, producing ~14 cepstral coefficients for each frame
- Calculate additional coefficients representing first- and second-order changes over time

Broadband spectrogram of speech

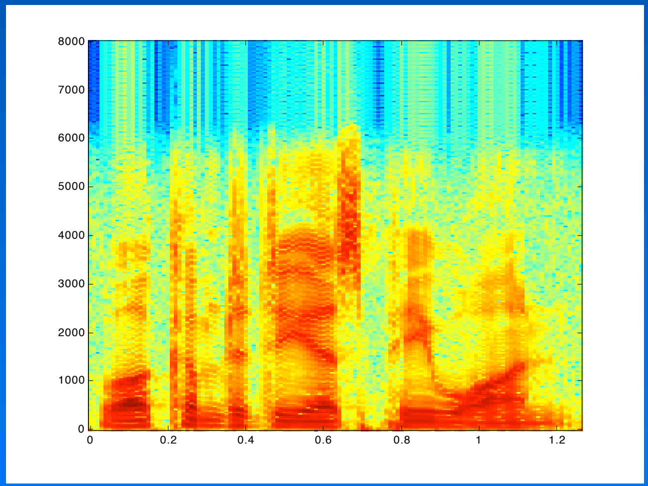


Cepstral representation

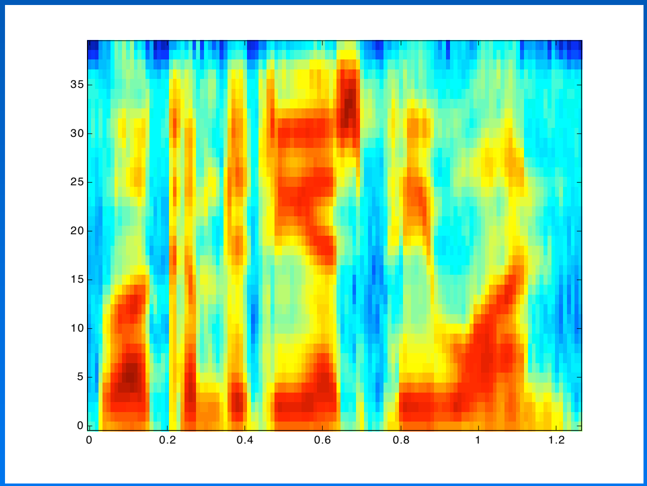


Comparing representations …

ORIGINAL SPEECH



CEPSTRAL REP

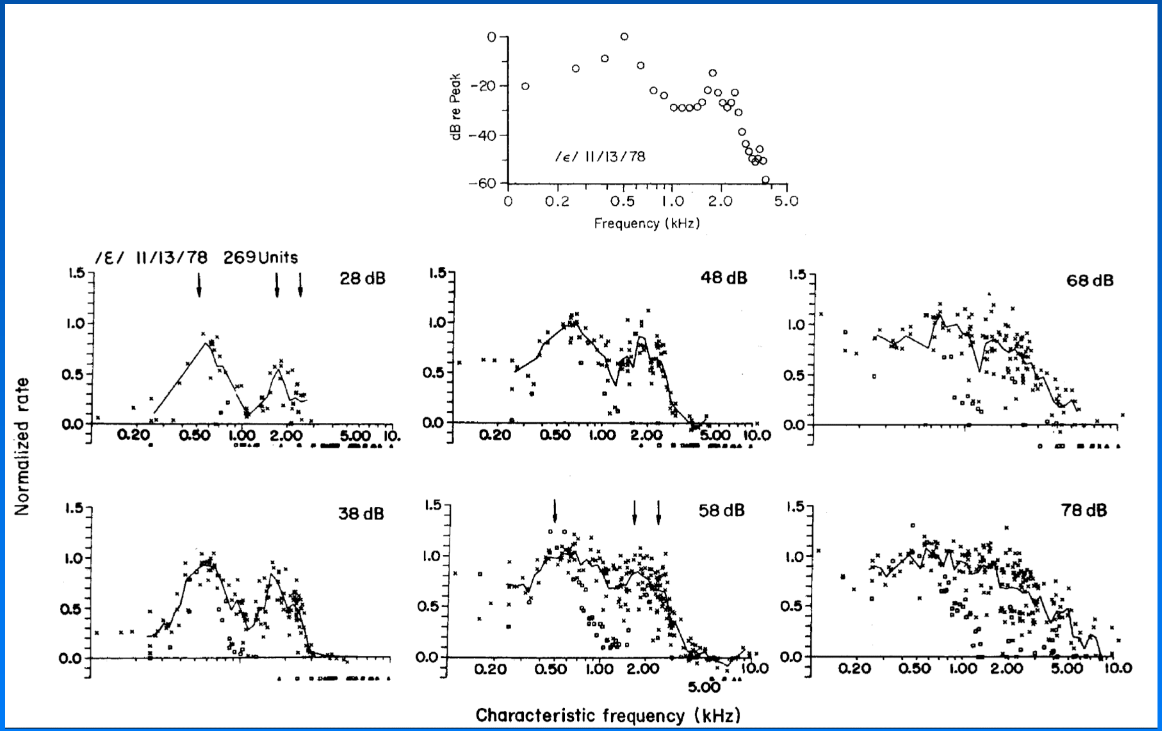


Comments on the MFCC representation

- It's very "blurry" compared to a wideband spectrogram!
- Aspects of auditory processing represented:
  - Frequency selectivity and spectral bandwidth (but using a constant analysis window duration!)
    - » Wavelet schemes exploit time-frequency resolution better
  - Nonlinear amplitude response (via log transformation only)
- Aspects of auditory processing NOT represented:
  - Detailed timing structure
  - Lateral suppression
  - Enhancement of temporal contrast
  - Other auditory nonlinearities

Speech representation using mean rate

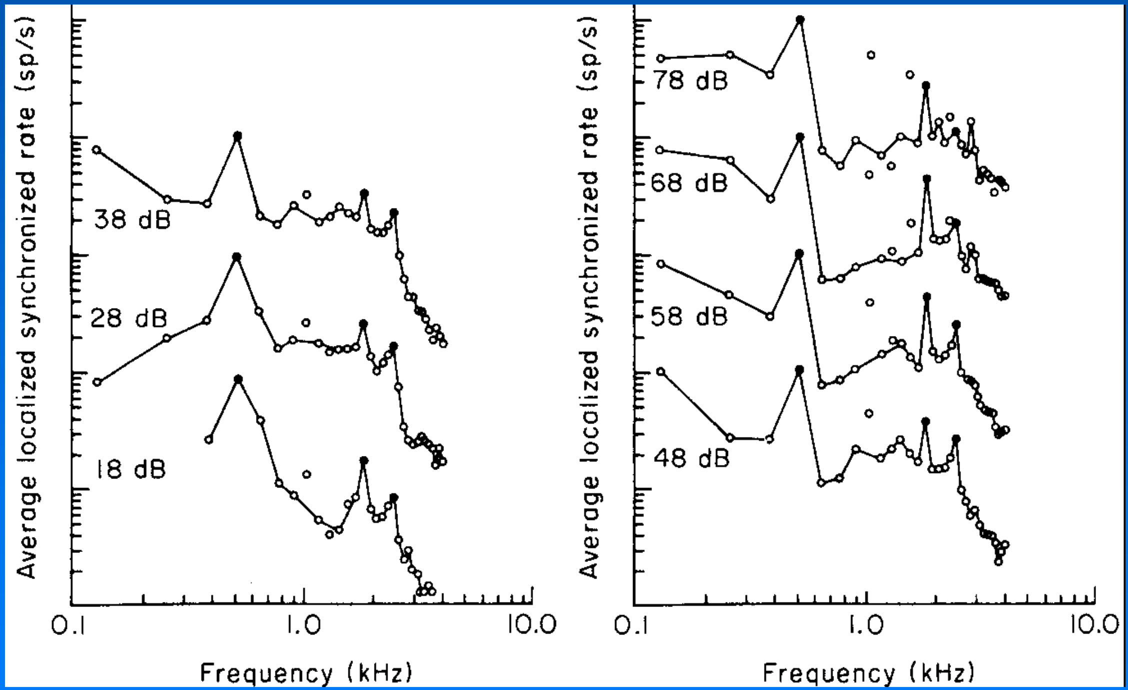
- Representation of vowels by Young and Sachs using mean rate:



- Mean rate representation does not preserve spectral information

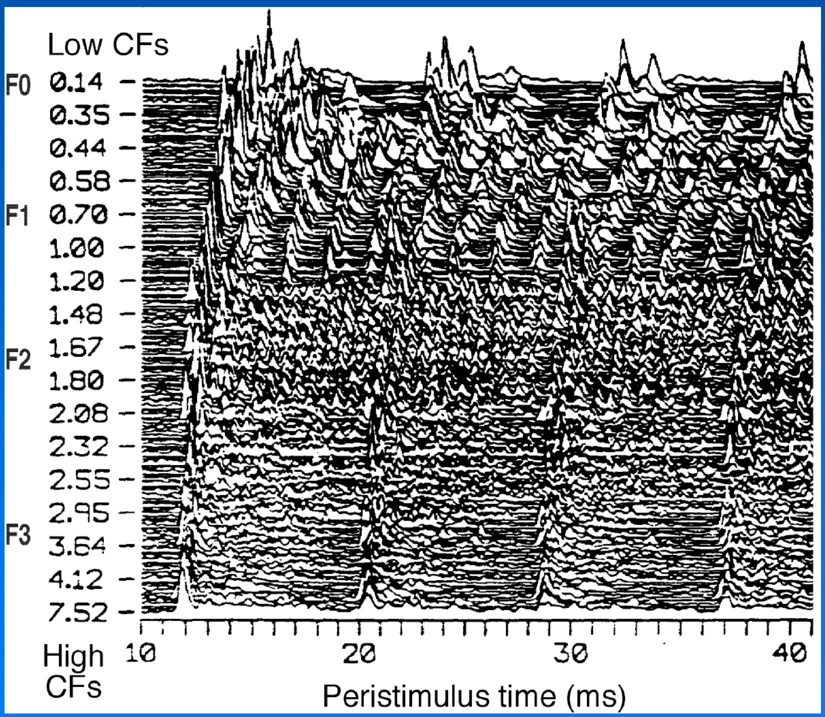
Speech representation using average localized synchrony measure

- Representation of vowels by Young and Sachs using ALSR:



The importance of timing information

- Re-analysis of Young-Sachs data by Searle:



- Temporal processing captures dominant formants in a spectral region

Paths to the realization of temporal fine structure in speech

- **Correlograms (Slaney and Lyon)**
- **Computations based on interval processing**
  - Seneff's Generalized Synchrony Detector (GSD) model
  - Ghitza's Ensemble Interval Histogram (EIH) model
  - D.C. Kim's Zero Crossing Peak Analysis (ZCPA) model

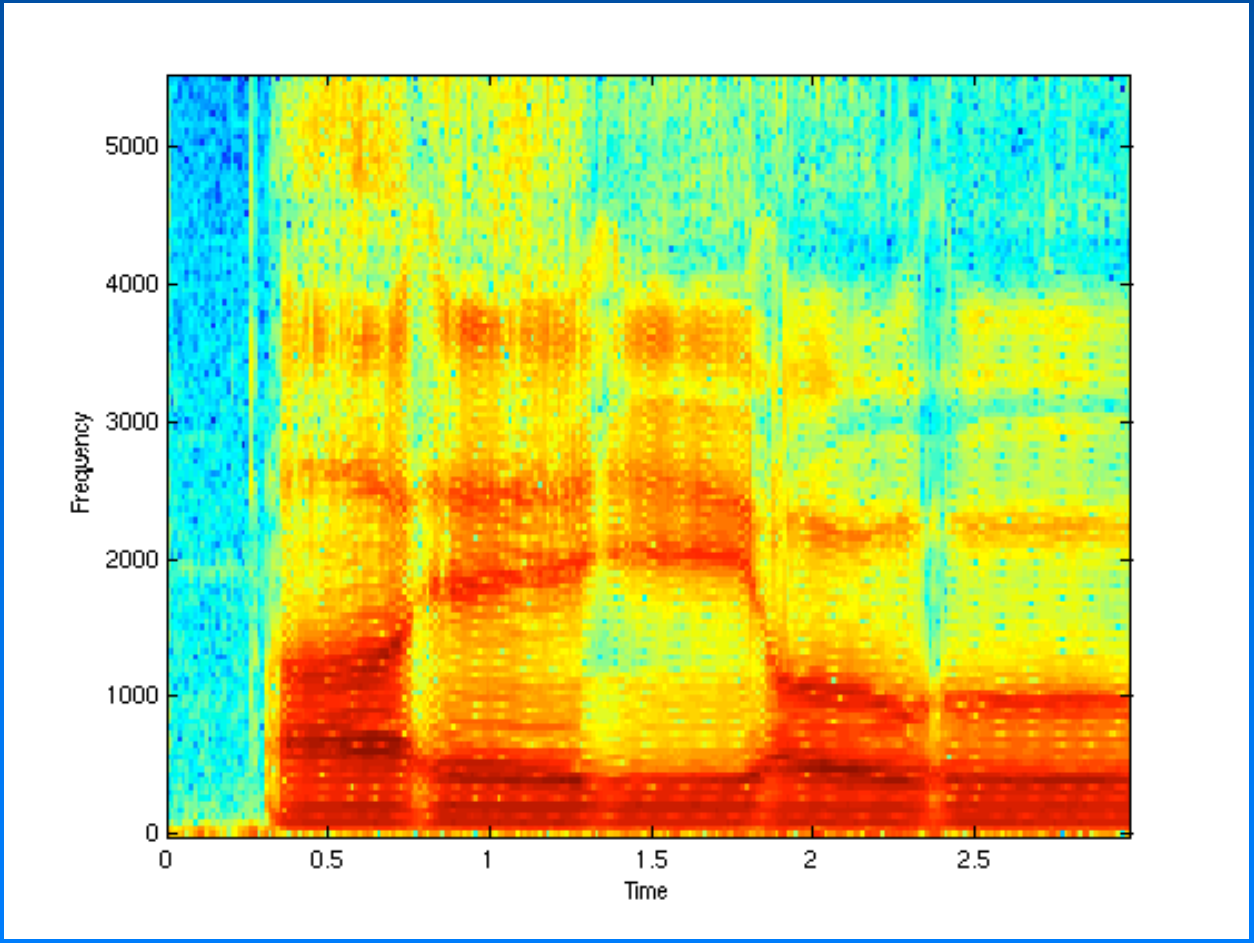
The original correlogram representation (Slaney and Lyon)

- **“Standard” peripheral auditory processing**
  – Bandpass filtering
  – Nonlinear rectification and compression
  – Other stuff
- **Autocorrelation of outputs of peripheral auditory model**
- **Analysis of 2-dimensional graph of autocorrelation vs CF, as it evolves over time**

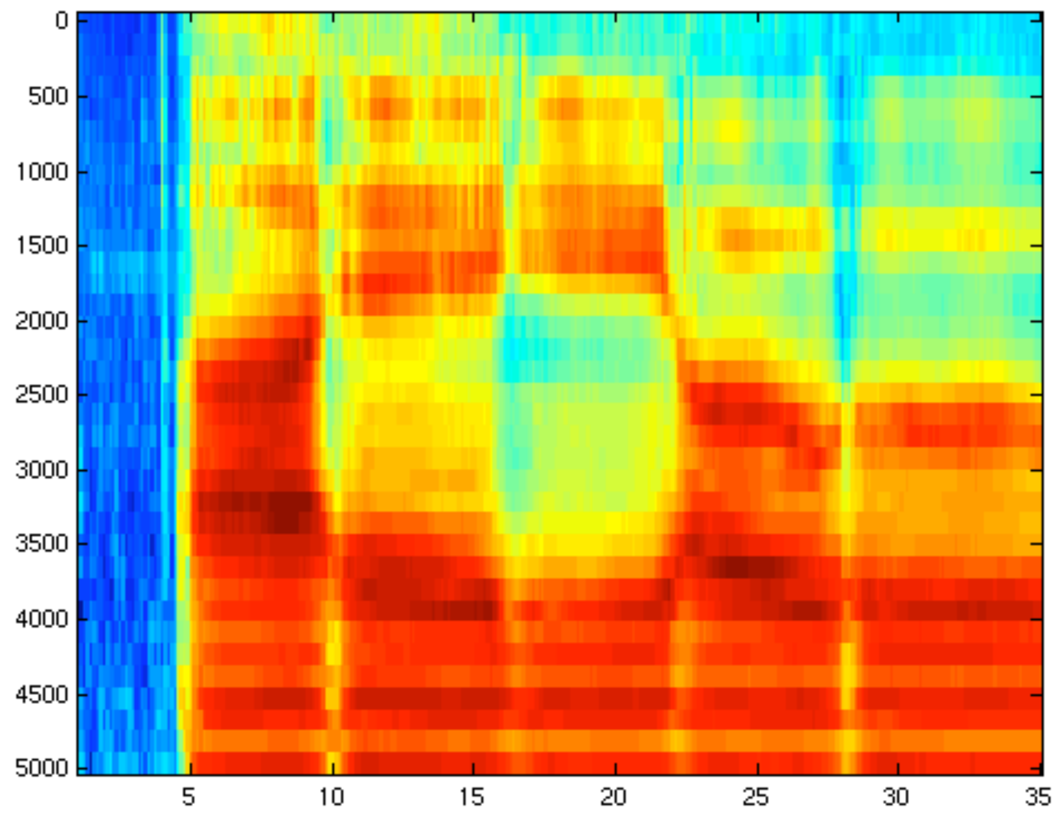
Modern timing-based representations

- **D. C. Kim's model:**
  - Bandpass filter
  - Extract zero crossings
  - Add frequency components in local regions based on inverses of times between zero crossings

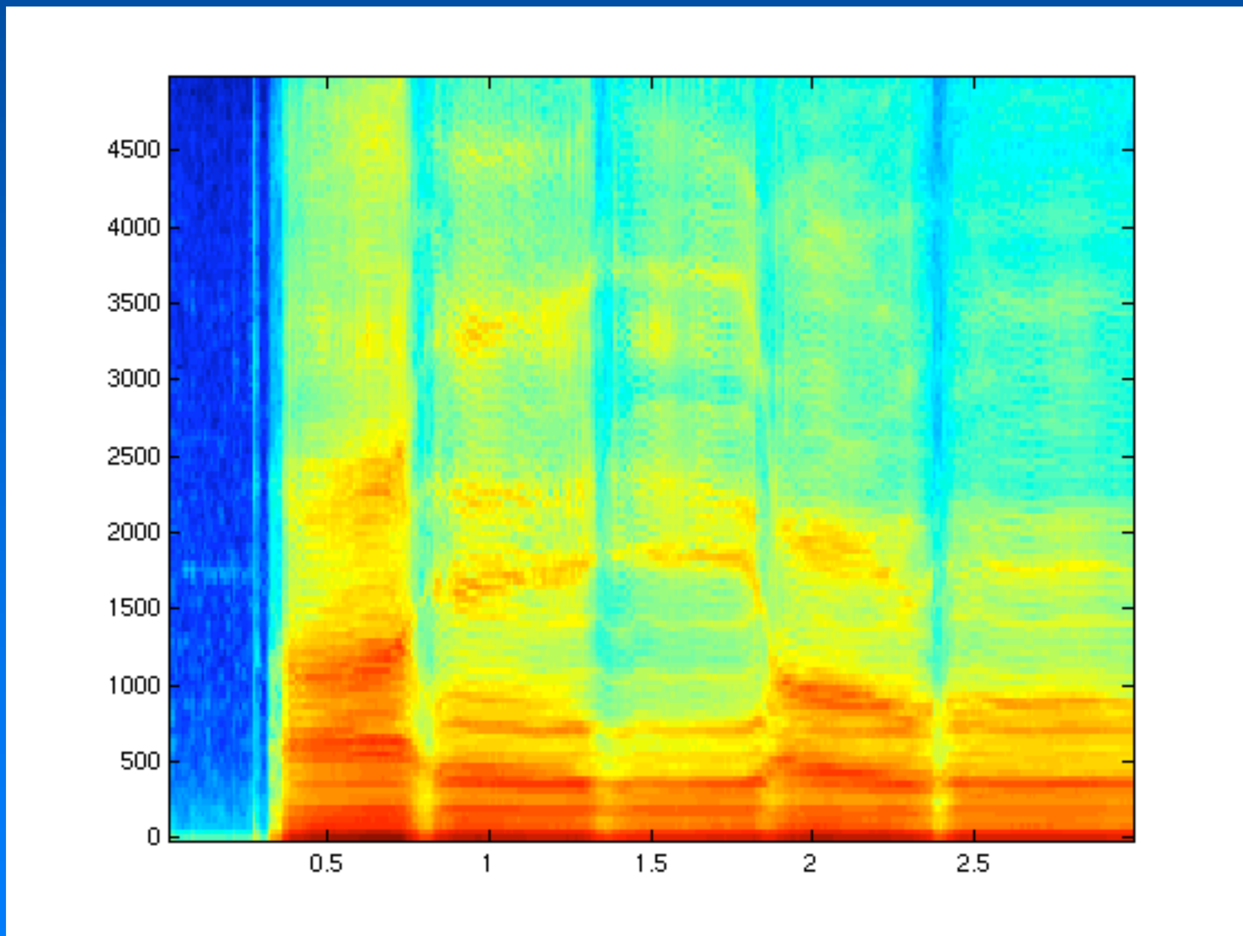
Another speech waveform



Vowels processed using energy only



Vowel sounds using autocorrelation expansion



Comments on peripheral timing information

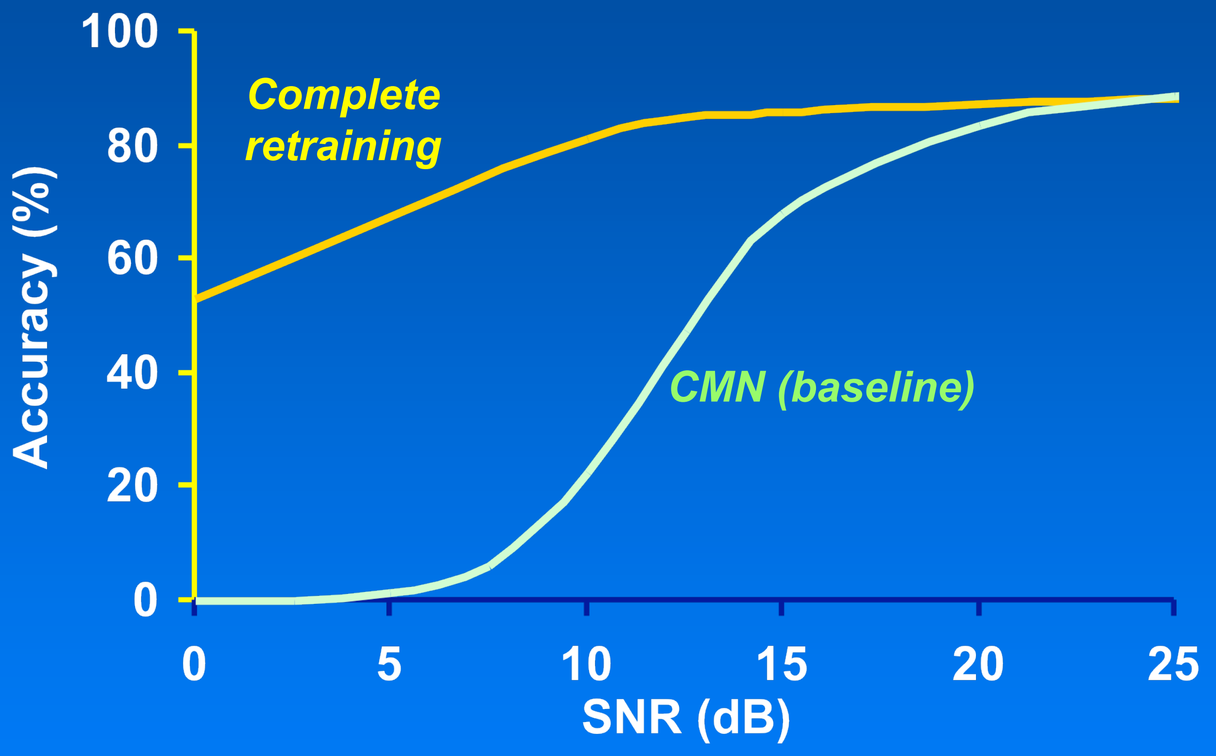
- Use of timing enables us to develop a rich display of frequencies, even with a limited number of analysis channels
- Nevertheless, this really gives us no new information unless the nonlinearities do something “interesting”
- Processing based on timing information (zero crossings, etc.) are likely to give us a more radically different display of info

Some of the hardest problems in speech recognition today ….

- Speech in high noise (Navy F-18 flight line)
- Speech in background speech
- Speech in background music
- Speech in reverberant environments

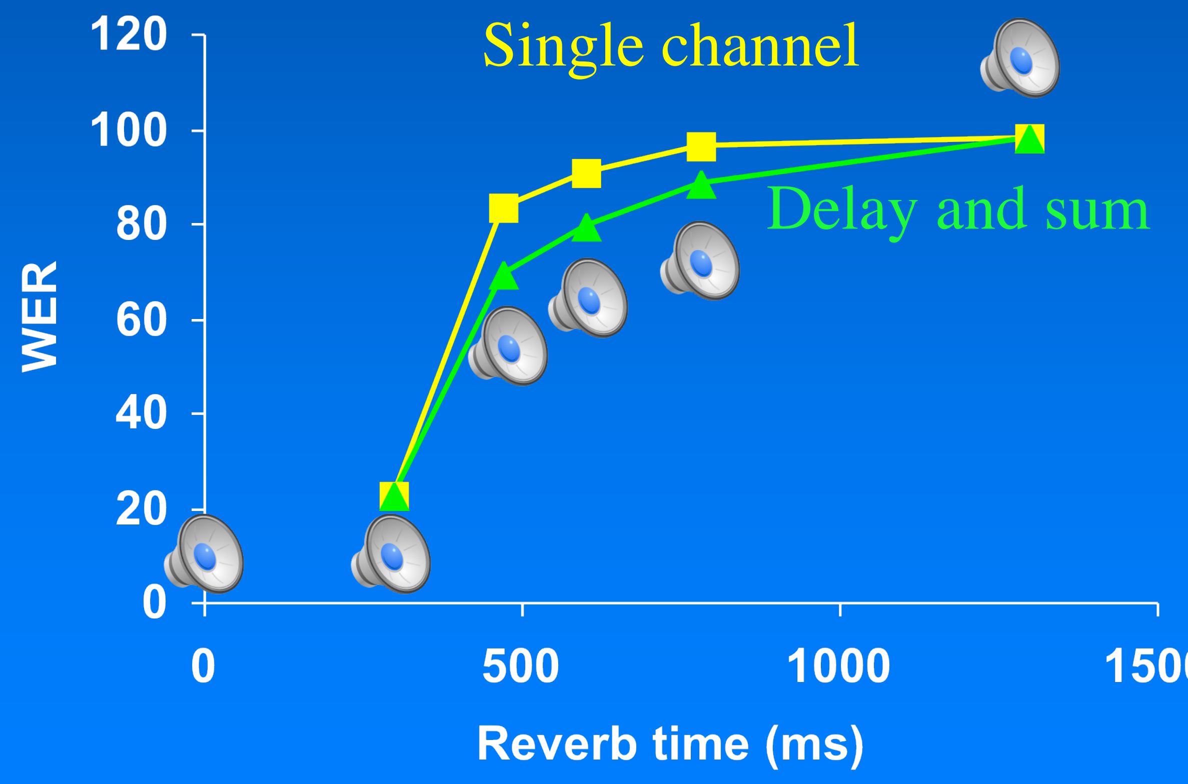
- Conventional signal processing provides only limited benefit for these problems

Speech recognition accuracy degrades in noise



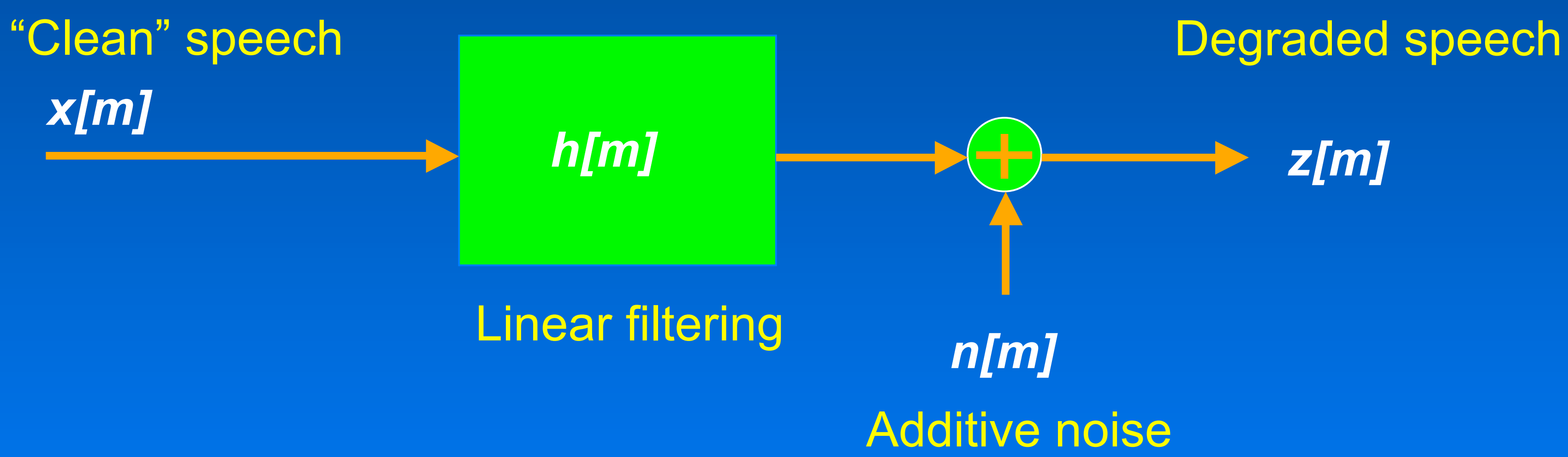
Recognition accuracy also degrades in highly reverberant rooms

- Comparison of single channel and delay-and-sum beamforming (WSJ data passed through measured impulse responses):



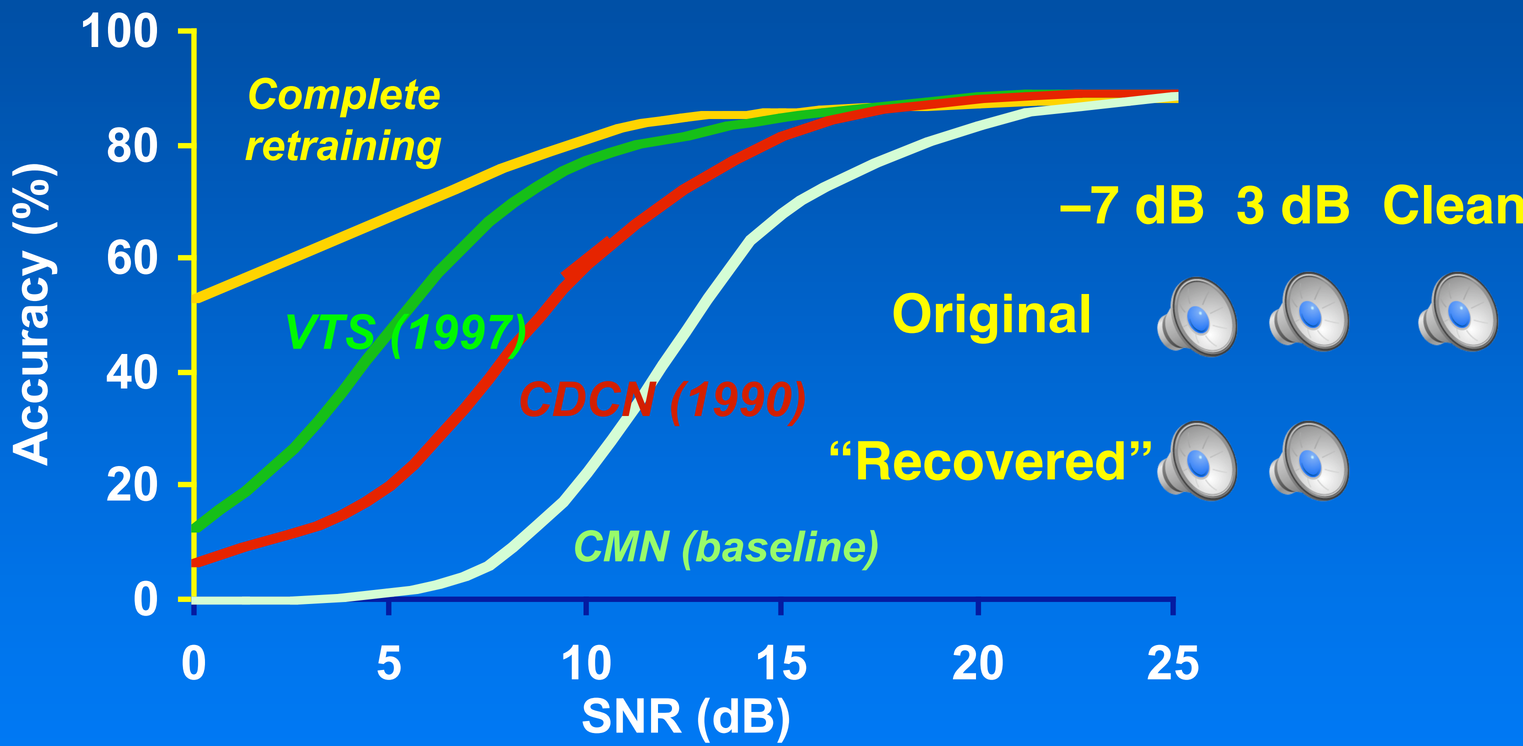
Carnegie Mellon

"Classical" solutions to robust speech recognition based on a model of the environment



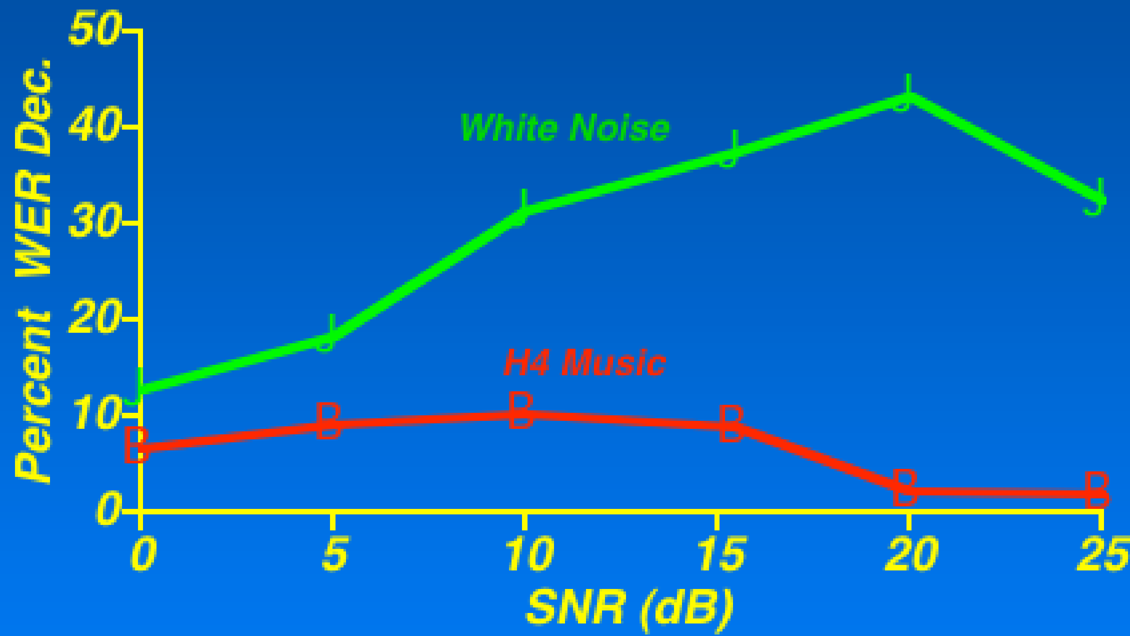
- **Compensation achieved by estimating parameters of noise and filter and applying inverse operations**

"Classical" compensation improves accuracy in stationary environments



- Threshold shifts by ~7 dB
- Accuracy still poor for low SNRs

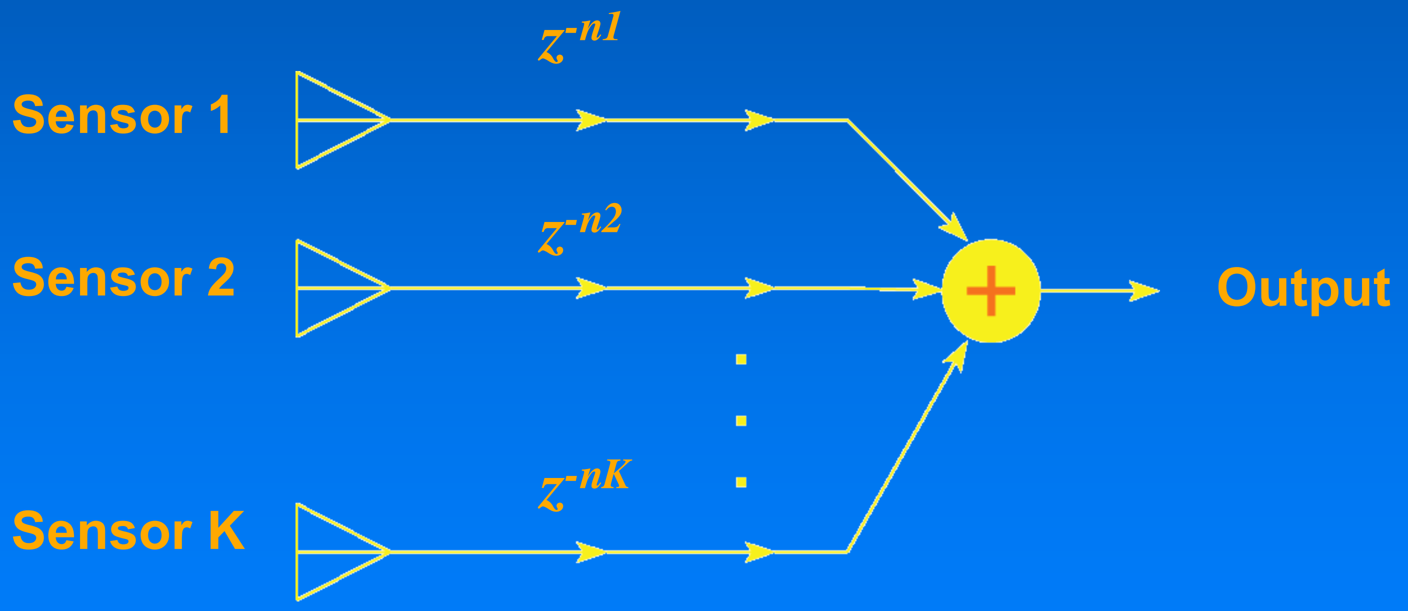
But model-based compensation does not improve accuracy (much) in transient noise



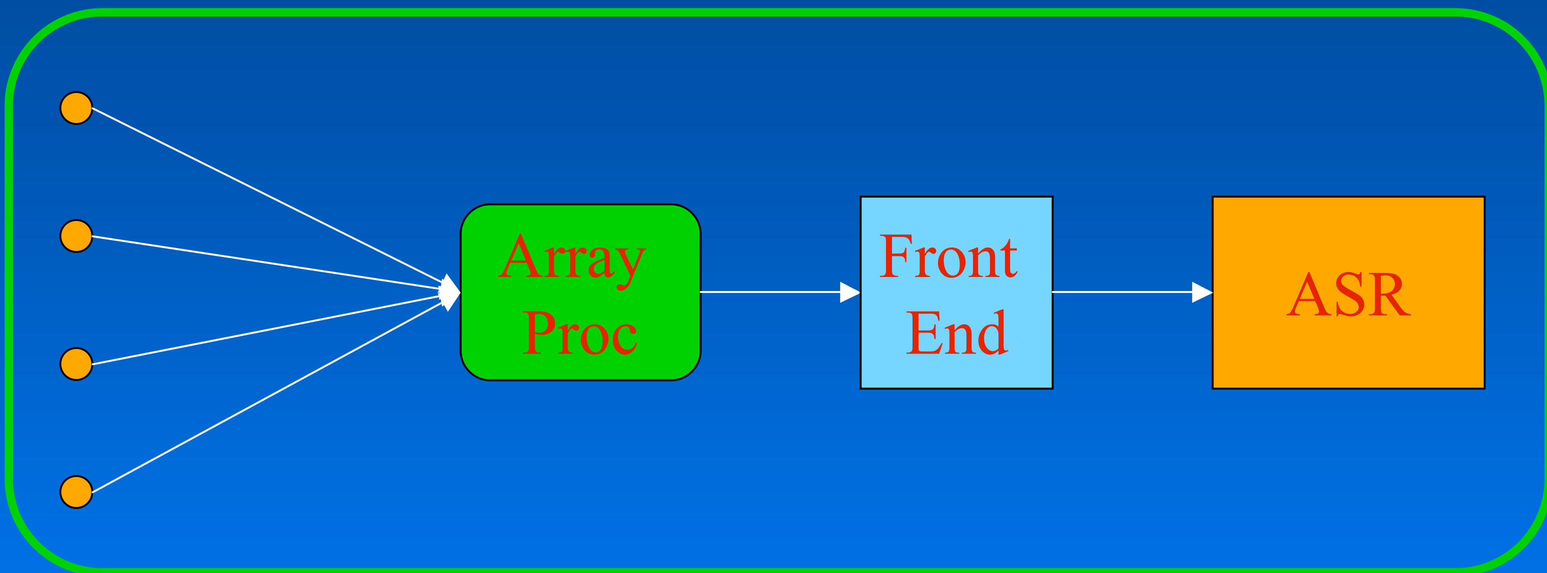
- Possible reasons: nonstationarity of background music and its speechlike nature

"Traditional" processing with multiple microphones: delay-and-sum beamforming

- Simple processing based on equalizing delays to sensors
- High directivity can be achieved with many sensors



Multi-microphone compensation for speech recognition based on cepstral distortion



- **Multi-microphone compensation based on optimizing speech features** rather than signal distortion (Seltzer '03)

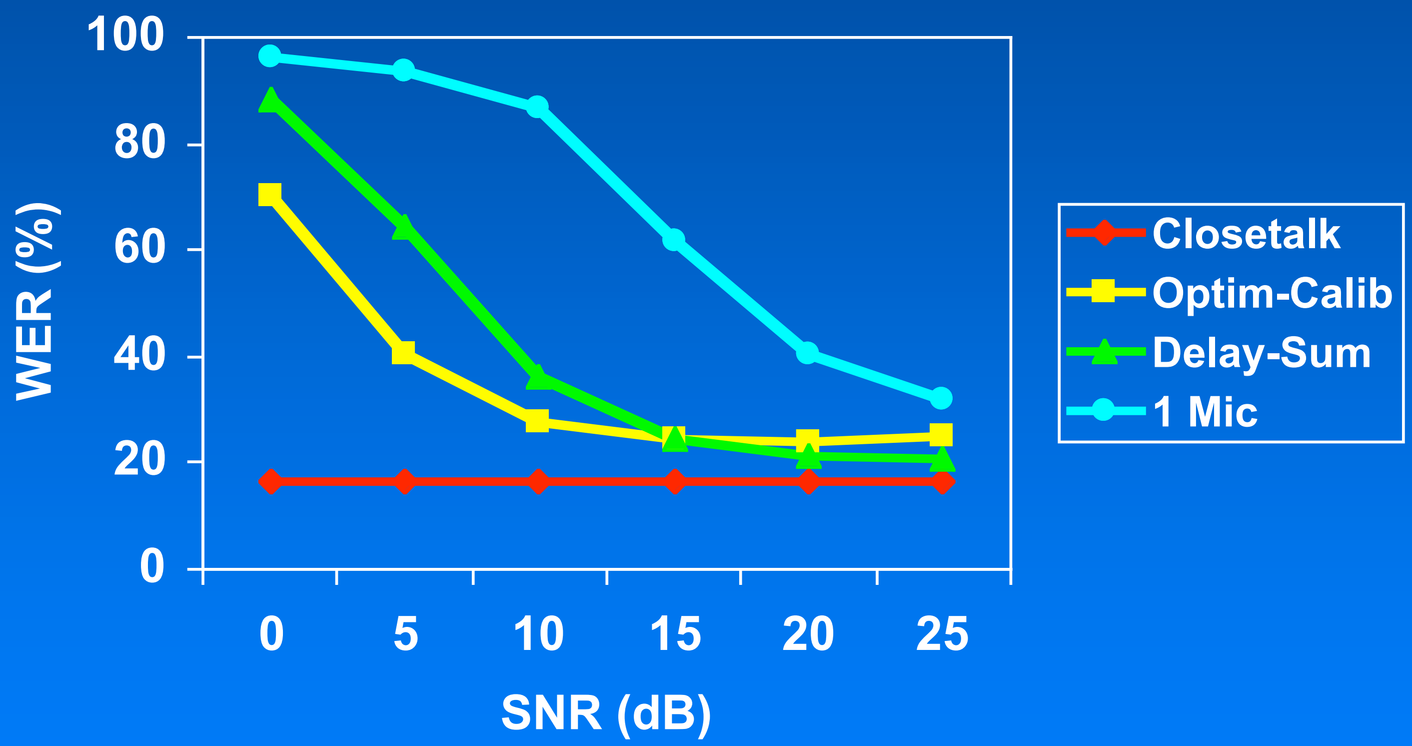
Speech in Room

Delay and Sum

Optimal Comp

Sample results using optimal array processing

- WER vs. SNR for WSJ with artificially-added white noise:

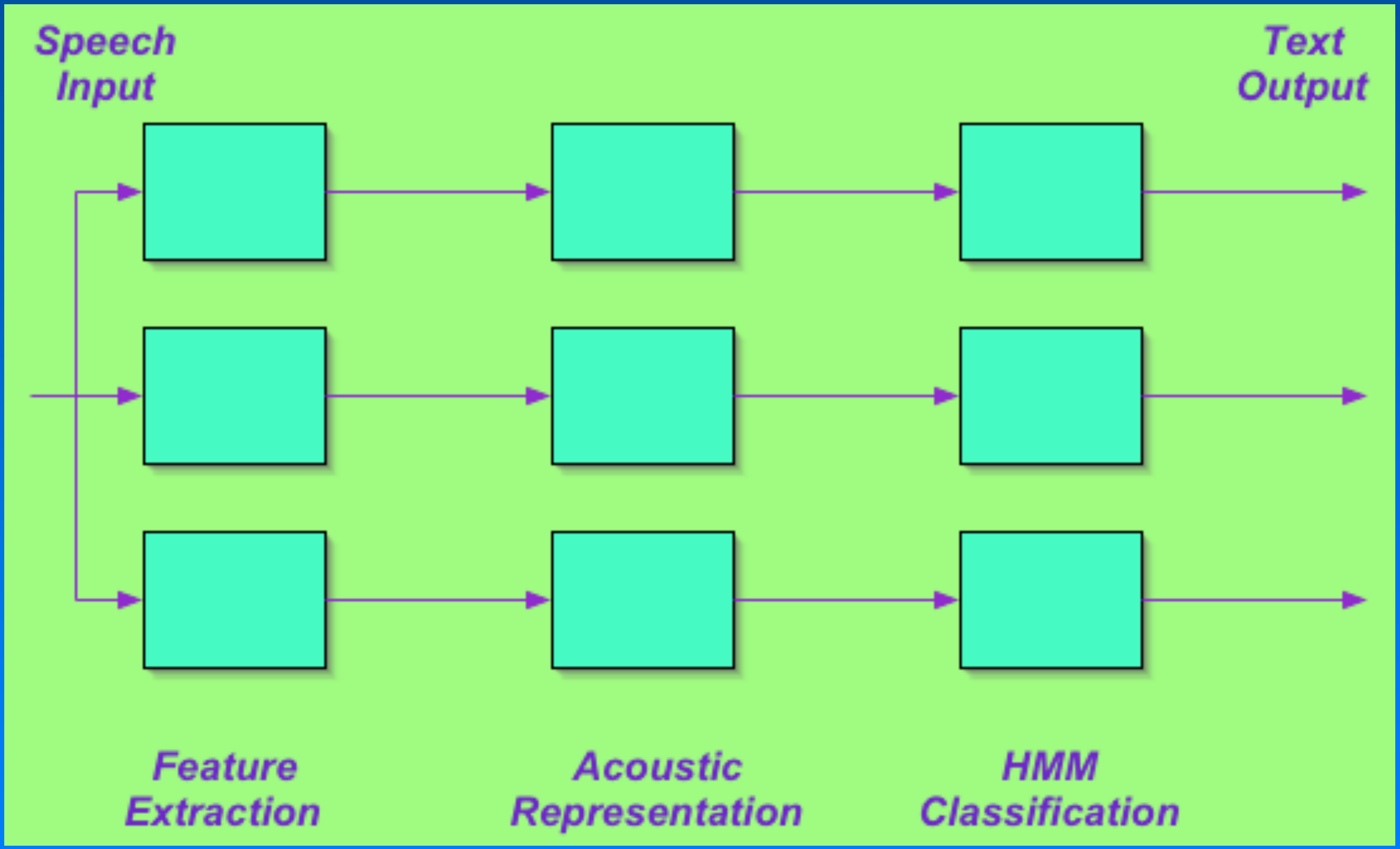


- Comment: Don’t trust results with artificially added noise!

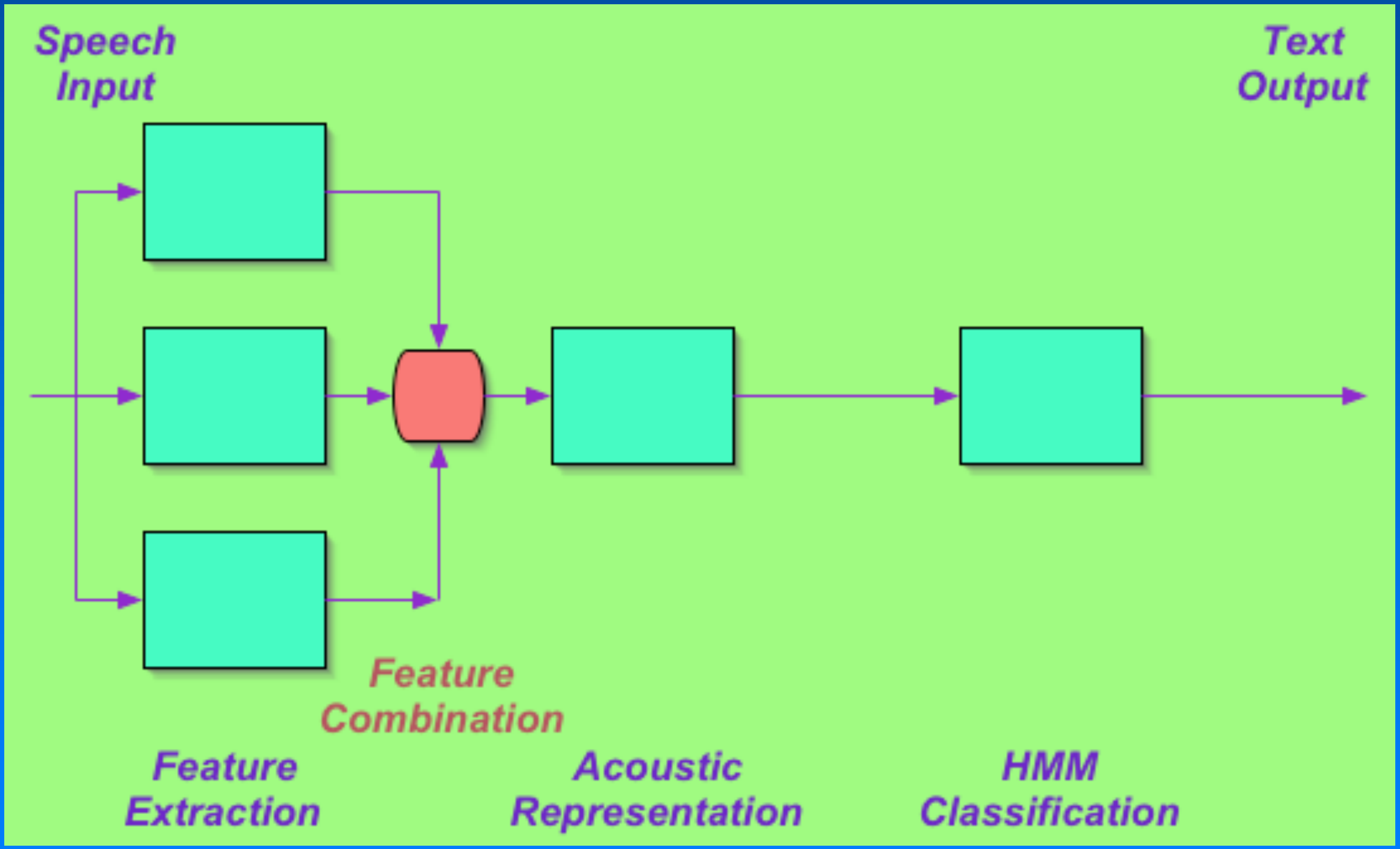
“Transitional” signal processing schemes: Multiband recognition and missing features

- **Multiband recognition** (***e.g.*** **Bourlard, Morgan, Hermansky *et al.***):
  - Decompose speech into several adjacent frequency bands
  - Train separate recognizers to process each band
  - Recombine information (somehow and somewhere)
- **Missing-feature recognition** (***e.g.*** **Cooke, Green, Raj *et al.***)
  - 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

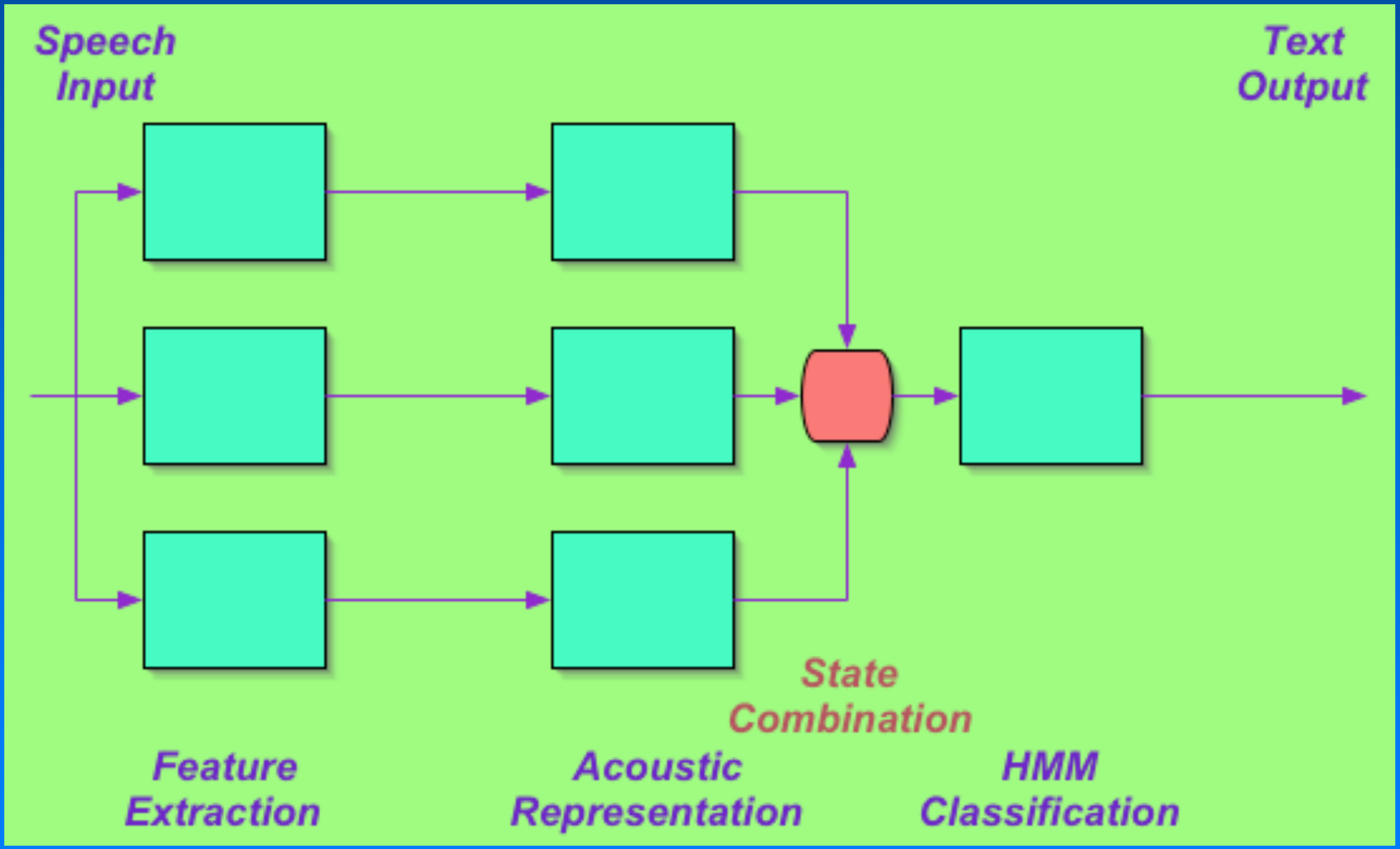
Combination of information streams: Independent recognition



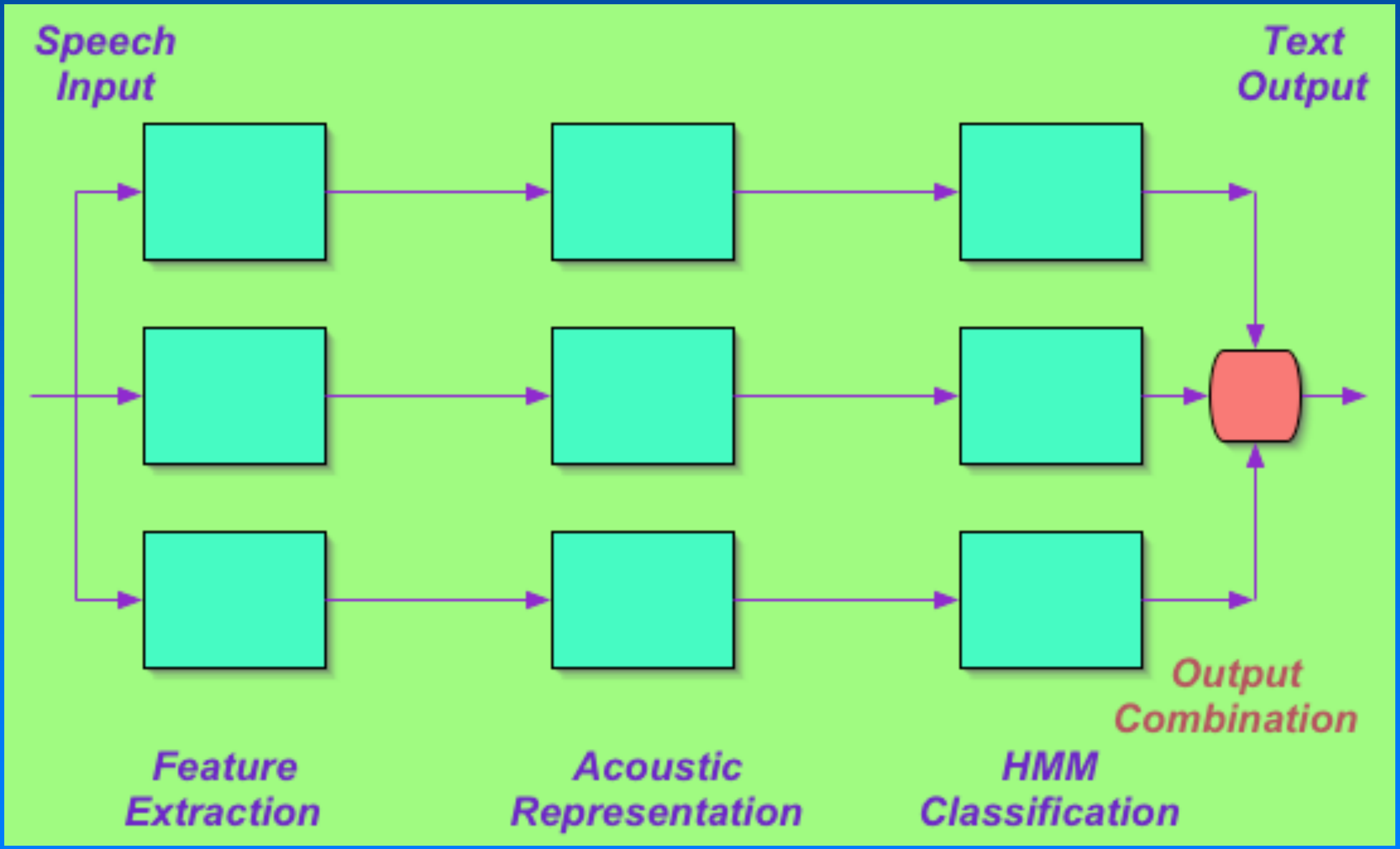
Combination of information streams: Feature combination



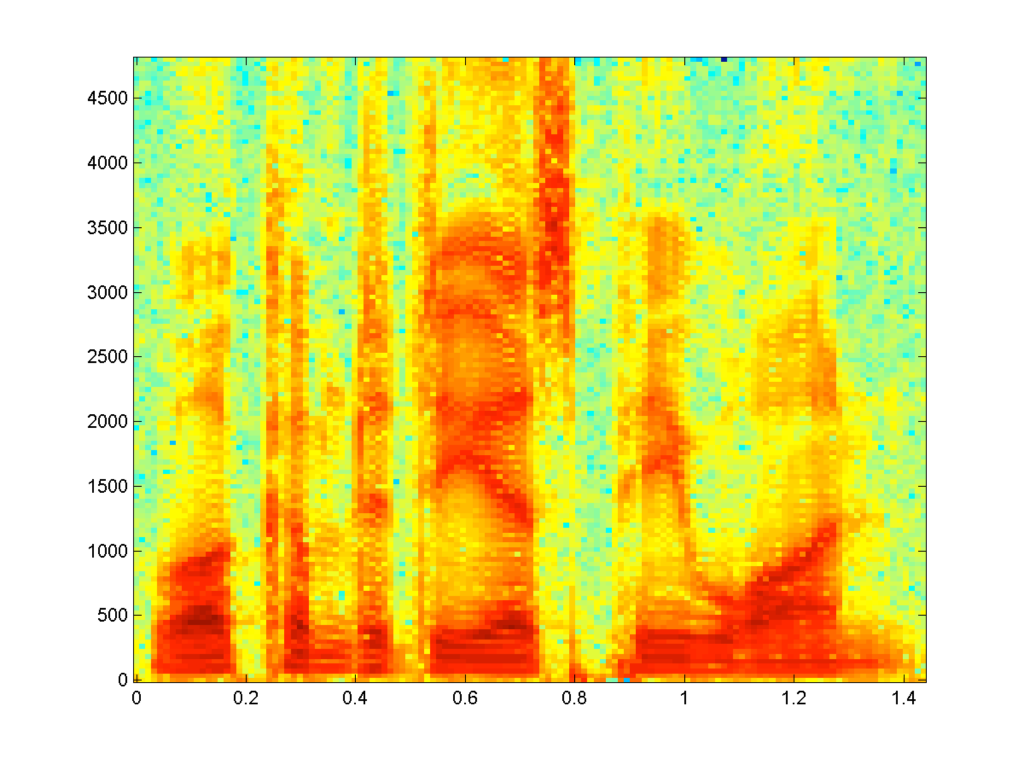
Combination of information streams: State/decoder combination



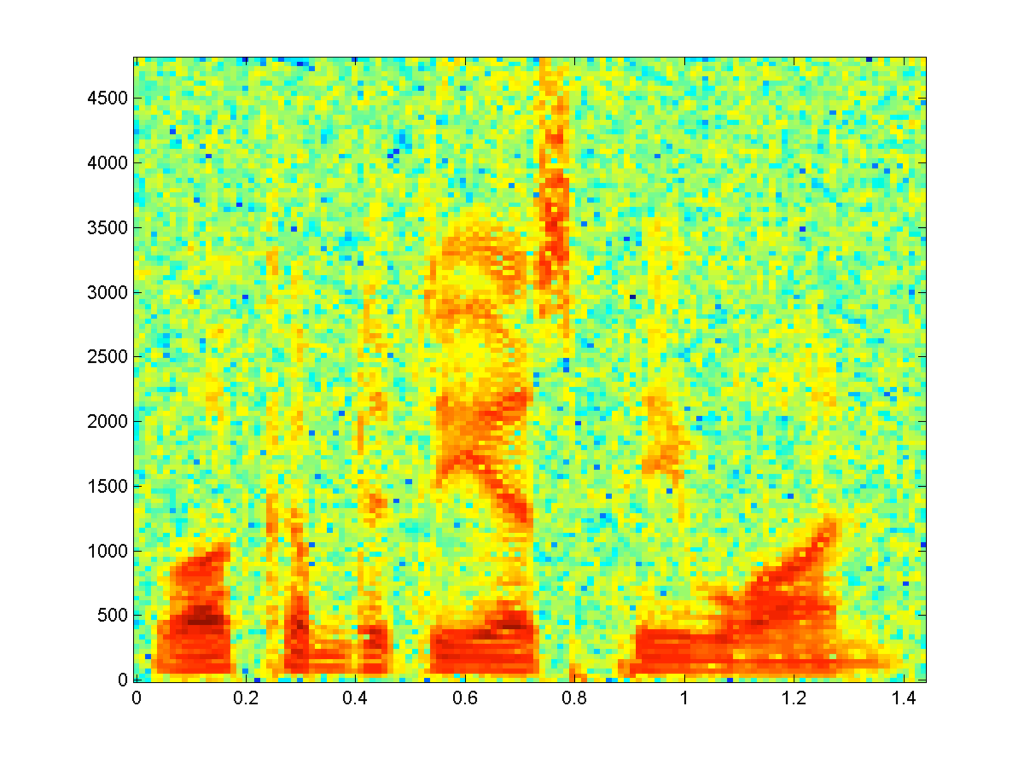
Combination of information streams: Output combination



Example of missing-feature analysis: an original speech spectrogram

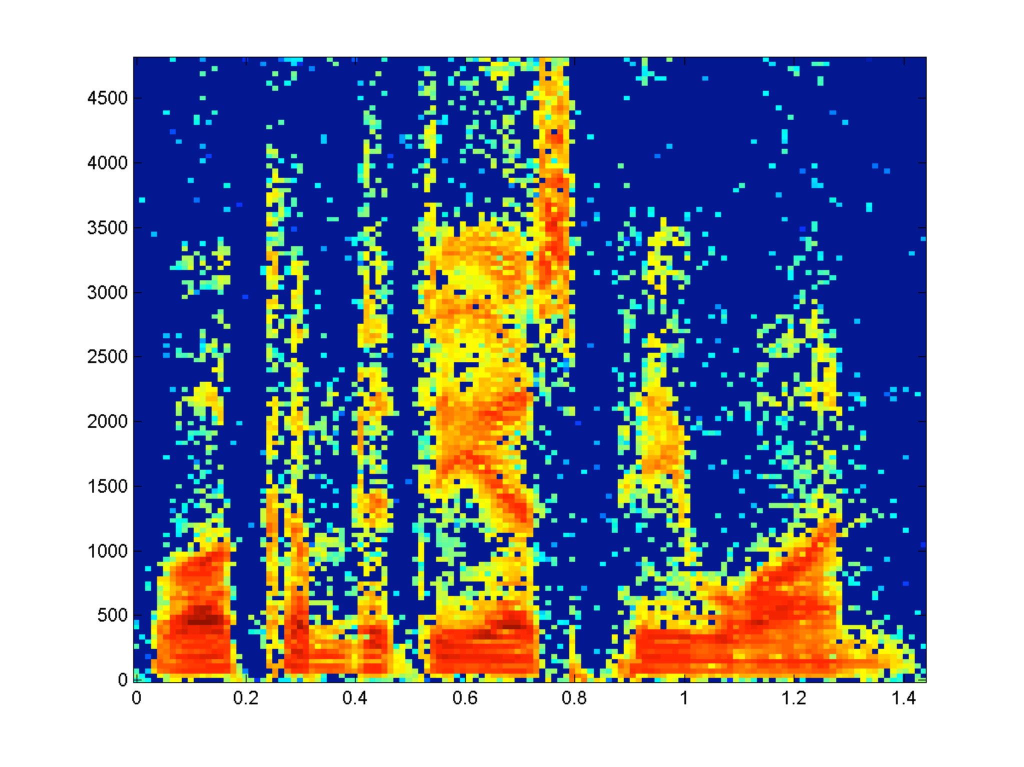


Spectrogram corrupted by noise at SNR 15 dB



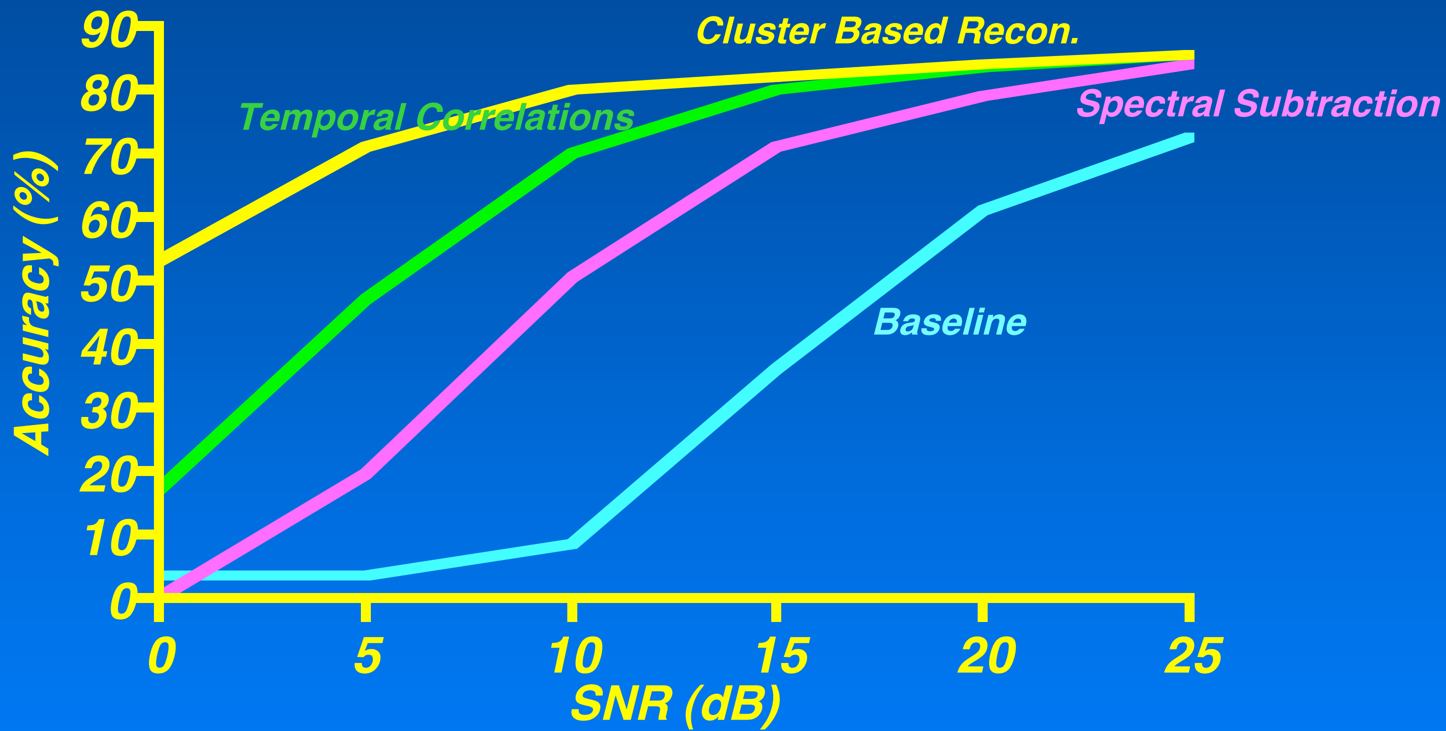
- Some regions are affected far more than others

Ignoring regions in the spectrogram that are corrupted by noise



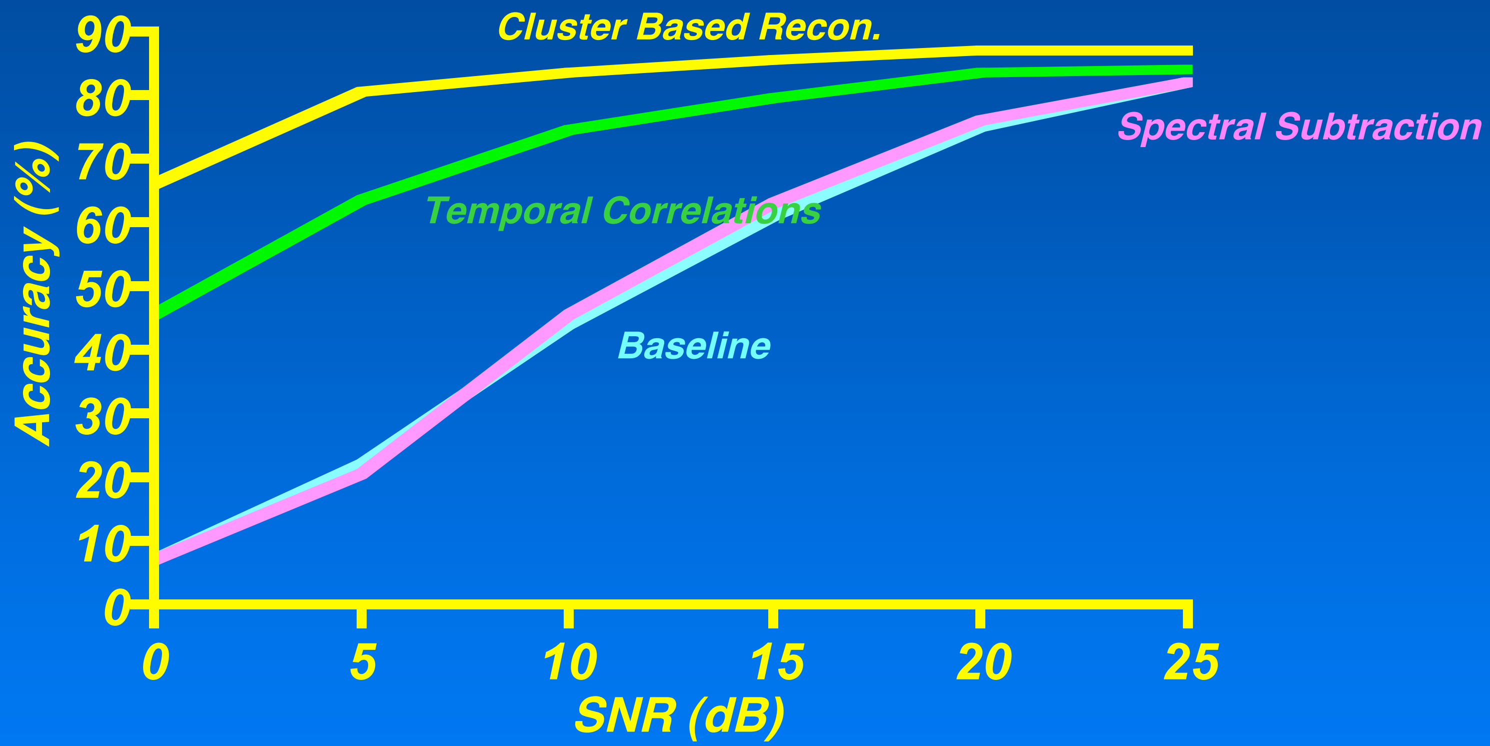
- All regions with SNR less than 0 dB deemed missing (dark blue)
- Recognition performed based on colored regions alone

Recognition accuracy using compensated cepstra, speech corrupted by white noise (Raj)



- **Caveat:** These results were obtained using **perfect knowledge** of missing feature "mask"
- **Big improvements** in SNR are possible

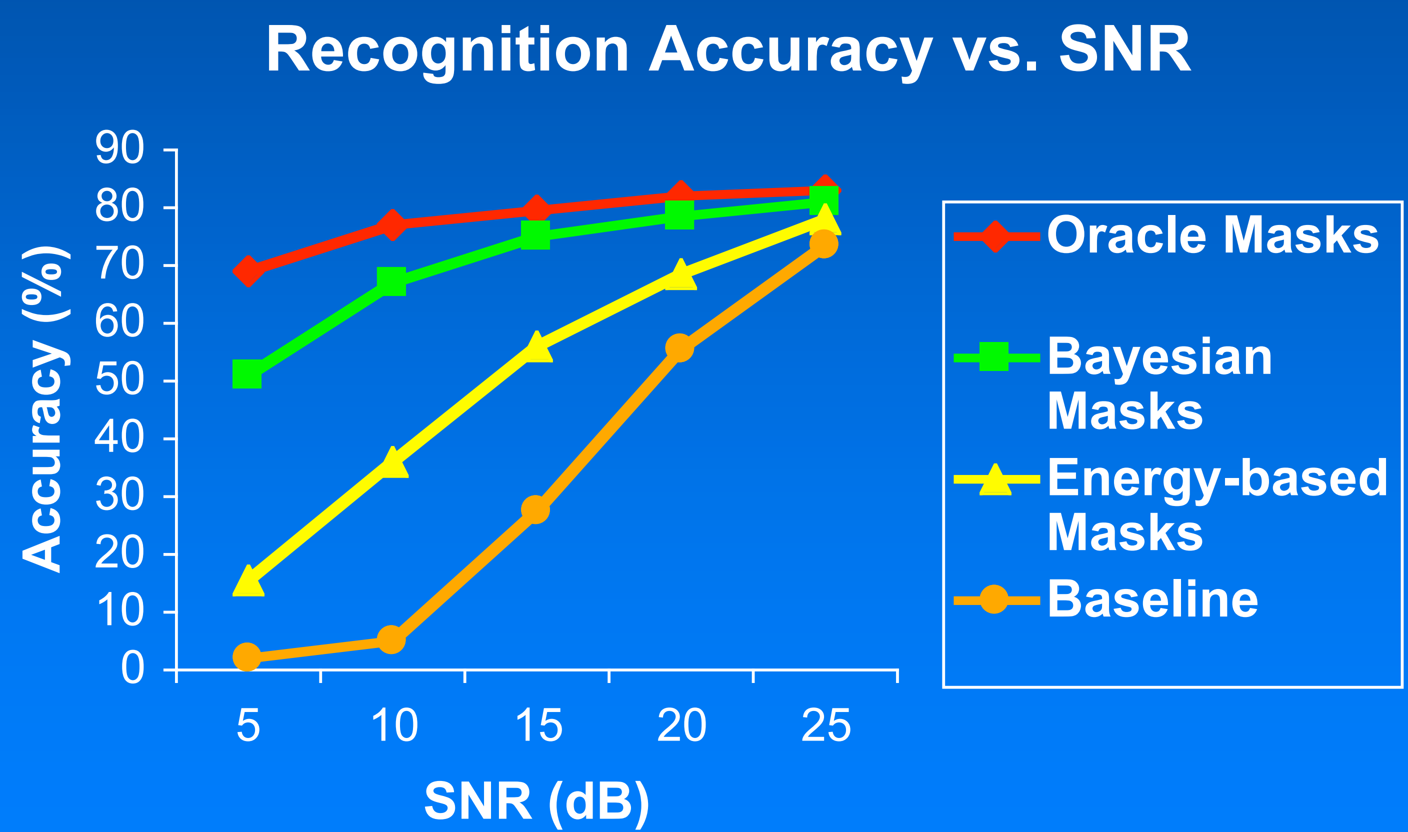
Recognition accuracy using compensated cepstra, speech corrupted by music



- 10-dB shift noted even for background music with ideal masks

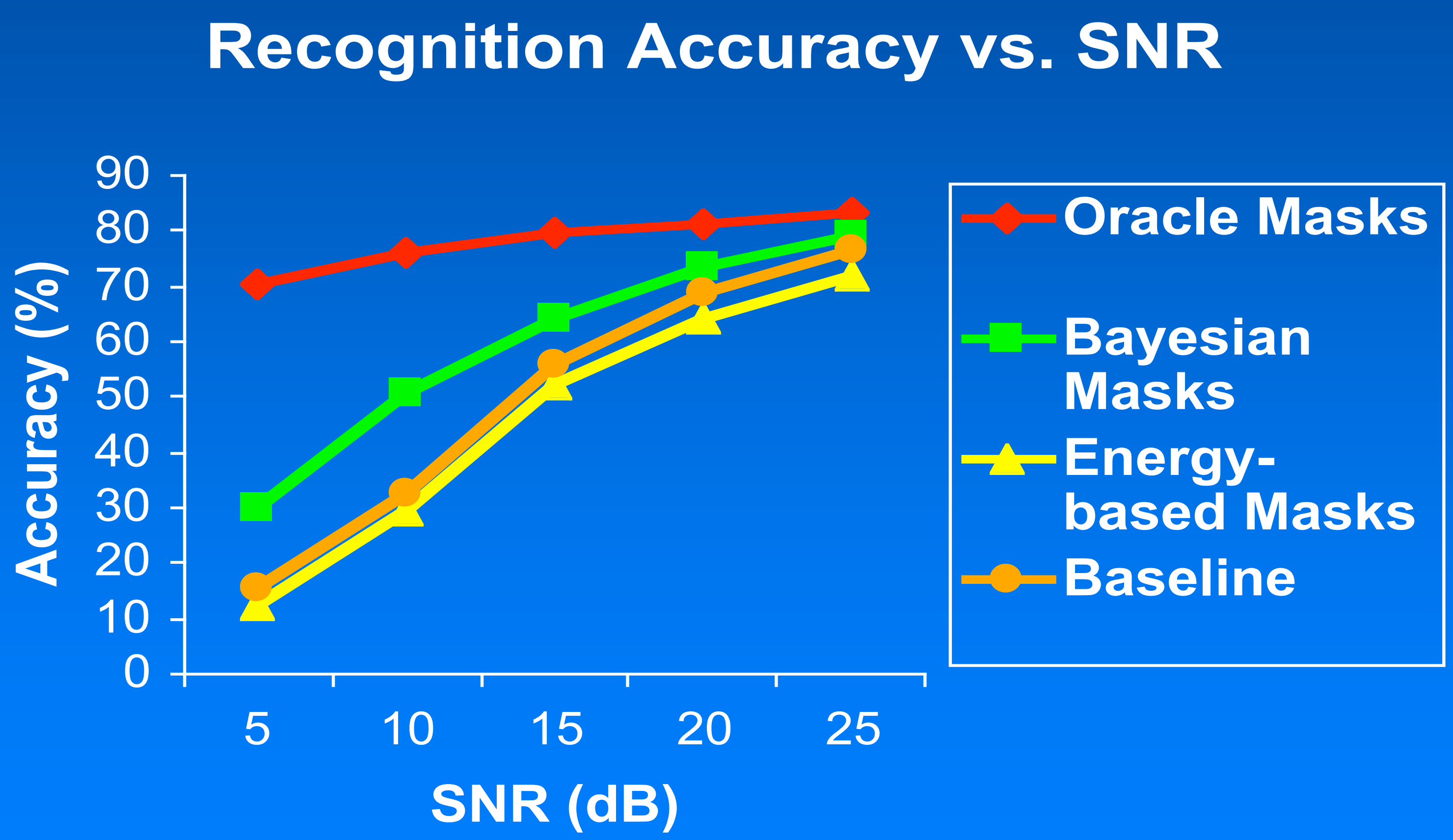
Practical recognition error using non-ideal masks: white noise (Seltzer)

- Speech plus White Noise:



Practical recognition error using non-ideal masks: background music

- Speech plus Music:



So what constitutes “modern” processing?

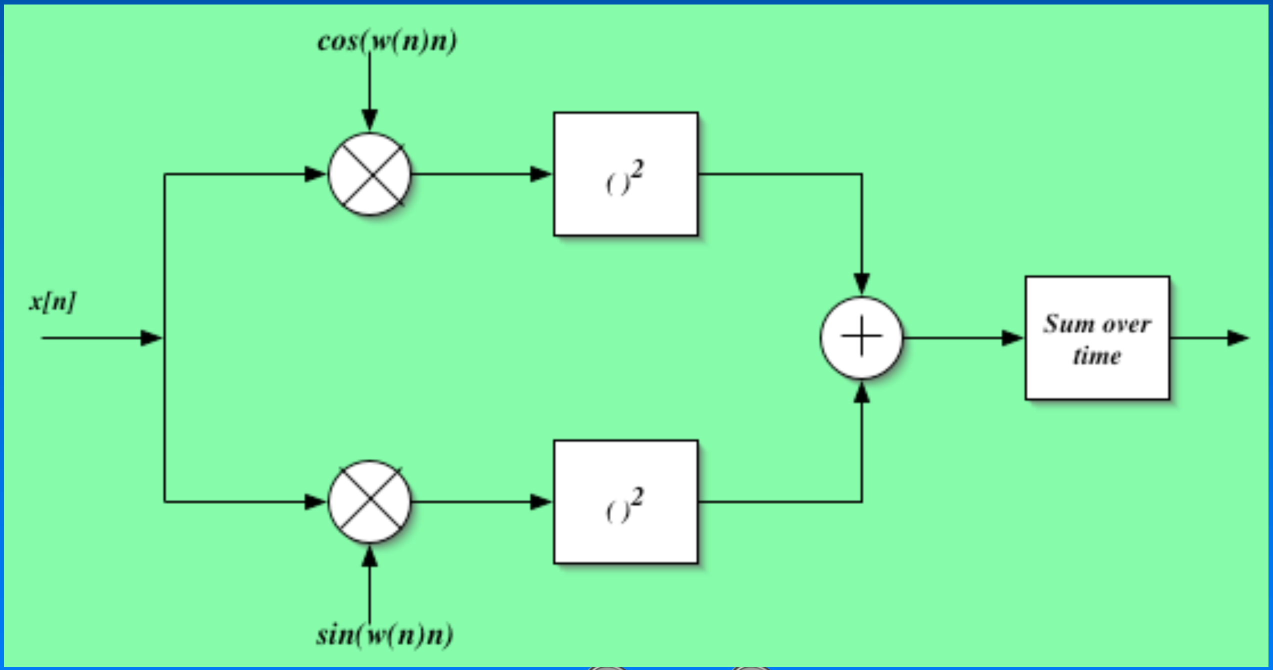
- Signal separation and robust recognition based on auditory scene analysis
- Signal separation and robust recognition based on better physiological and perceptual models

Using auditory streaming cues to separate sound sources

- **Many groups are now working to extract cues identified by Al Bregman and his colleagues to separate and group auditory fragments that are believed to arise from different sources**
- **Most commonly-discussed cues:**
  – Fundamental frequency/harmonicity
  – Source location/interaural time delay (ITD)/interaural correlation
- **Other cues that have been studied:**
  – Frequency and amplitude modulation
  – Common onset and offset
- **Comment:** Results so far are just the tip of the iceberg

One pitch-based approach: synchronized heterodyne analysis

- Extract instantaneous pitch, extract amplitudes at harmonics, resynthesize



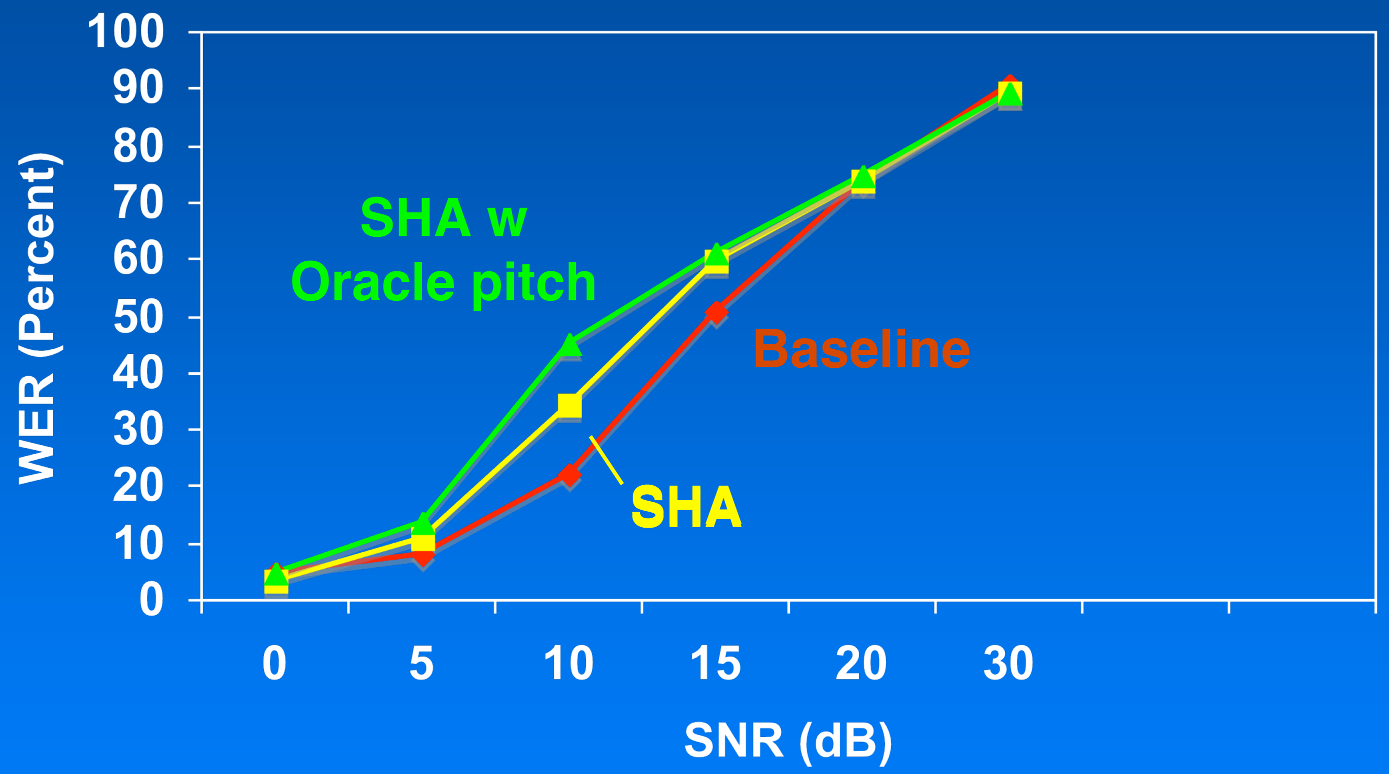
- Original speech samples:
- Reconstructed speech:

Separating speech signals by heterodyne analysis

- Combined speech signals:
- Speech separated by heterodyne filters:
- Comment: men mask women more because upper male harmonics are more likely to impinge on lower female harmonics

Carnegie Mellon

Speech recognition in noise based on pitch tracking



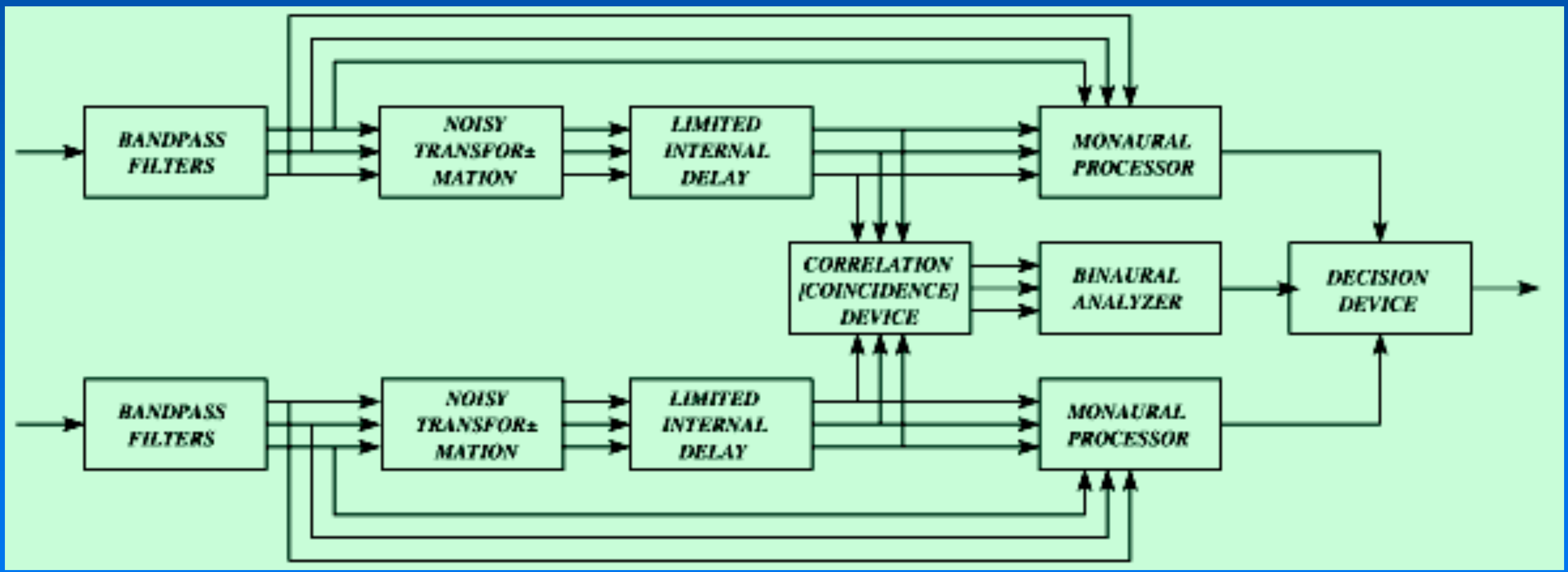
- Initial results could improve as techniques mature

Speech separation by source location

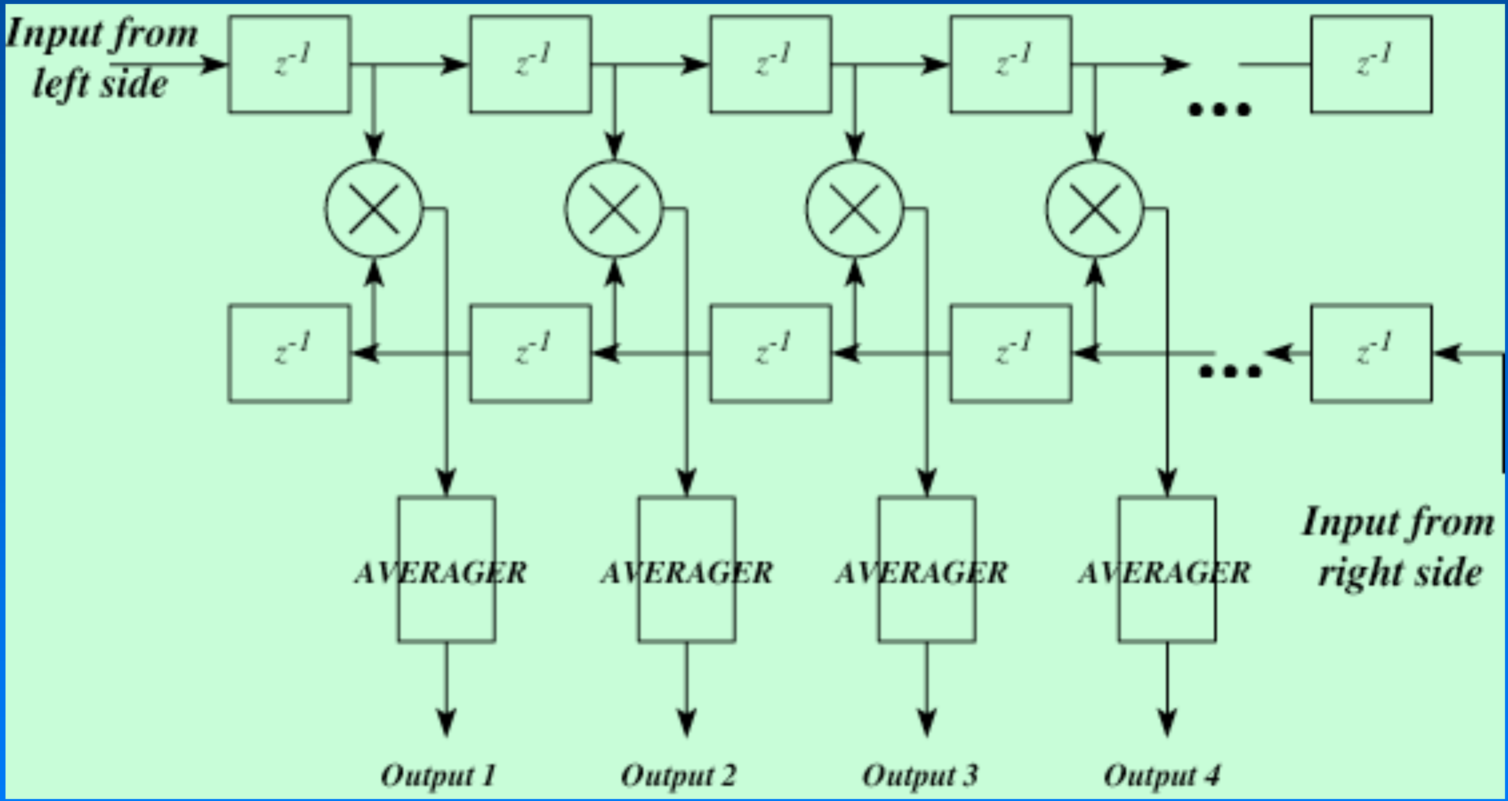
- **Sources arriving from different azimuths produce interaural time delays (ITDs) and interaural intensity differences (IIDs) as they arrive at the two ears**
- **So far this information has been used for**
  - Better “masks” for missing feature recognition and to combat reverberation (*e.g.* Brown*,* Wang *et al.*)
  - Direct separation from interaural representation

Carnegie Mellon

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



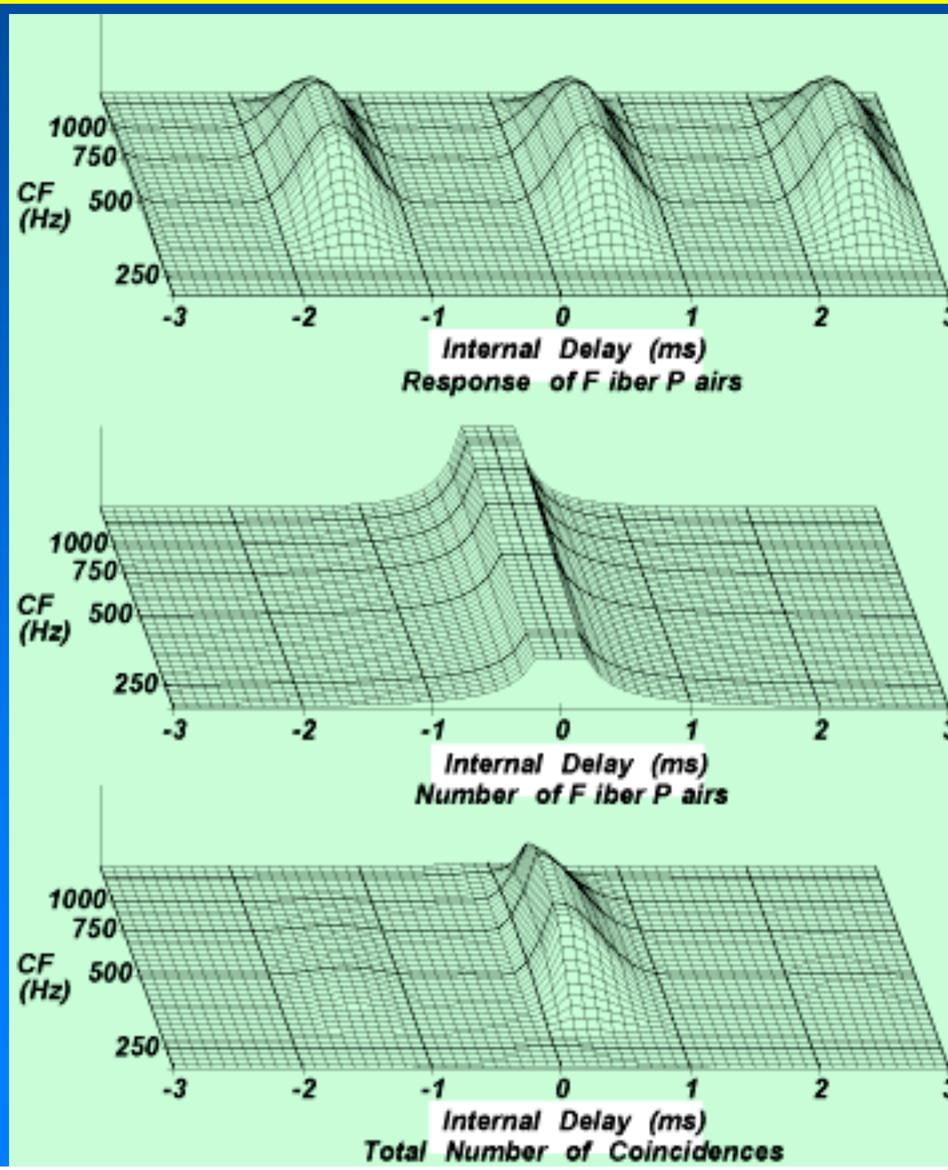
Jeffress's model of ITD extraction (1948)



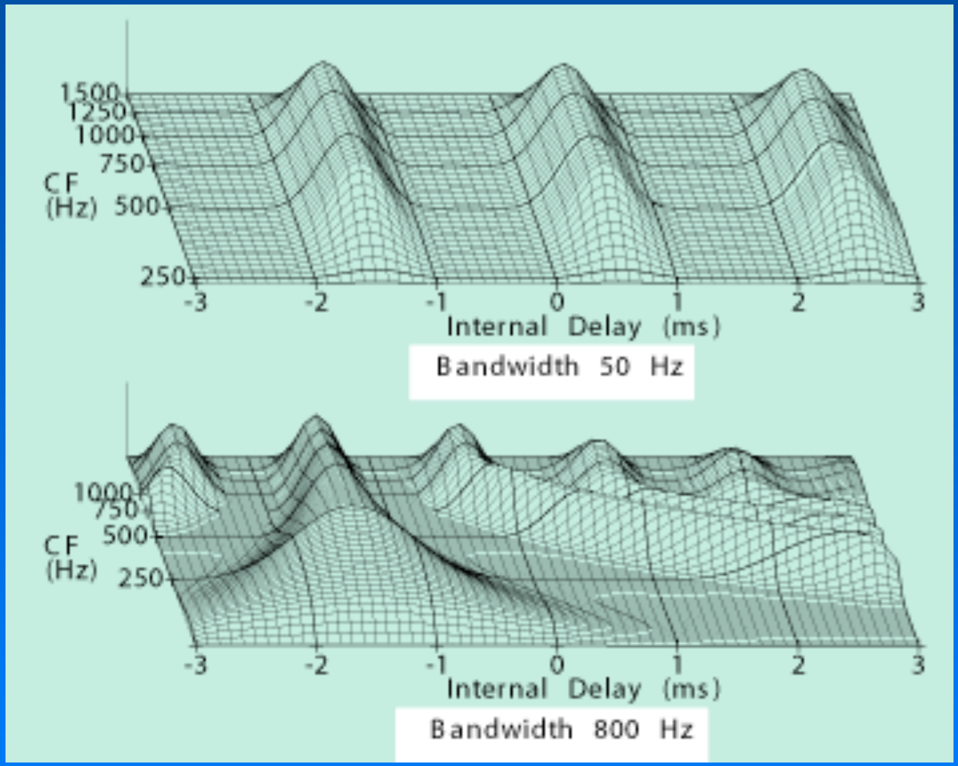
Comment: Several alternates have been proposed recently for correlation-based mechanism

Carnegie Mellon

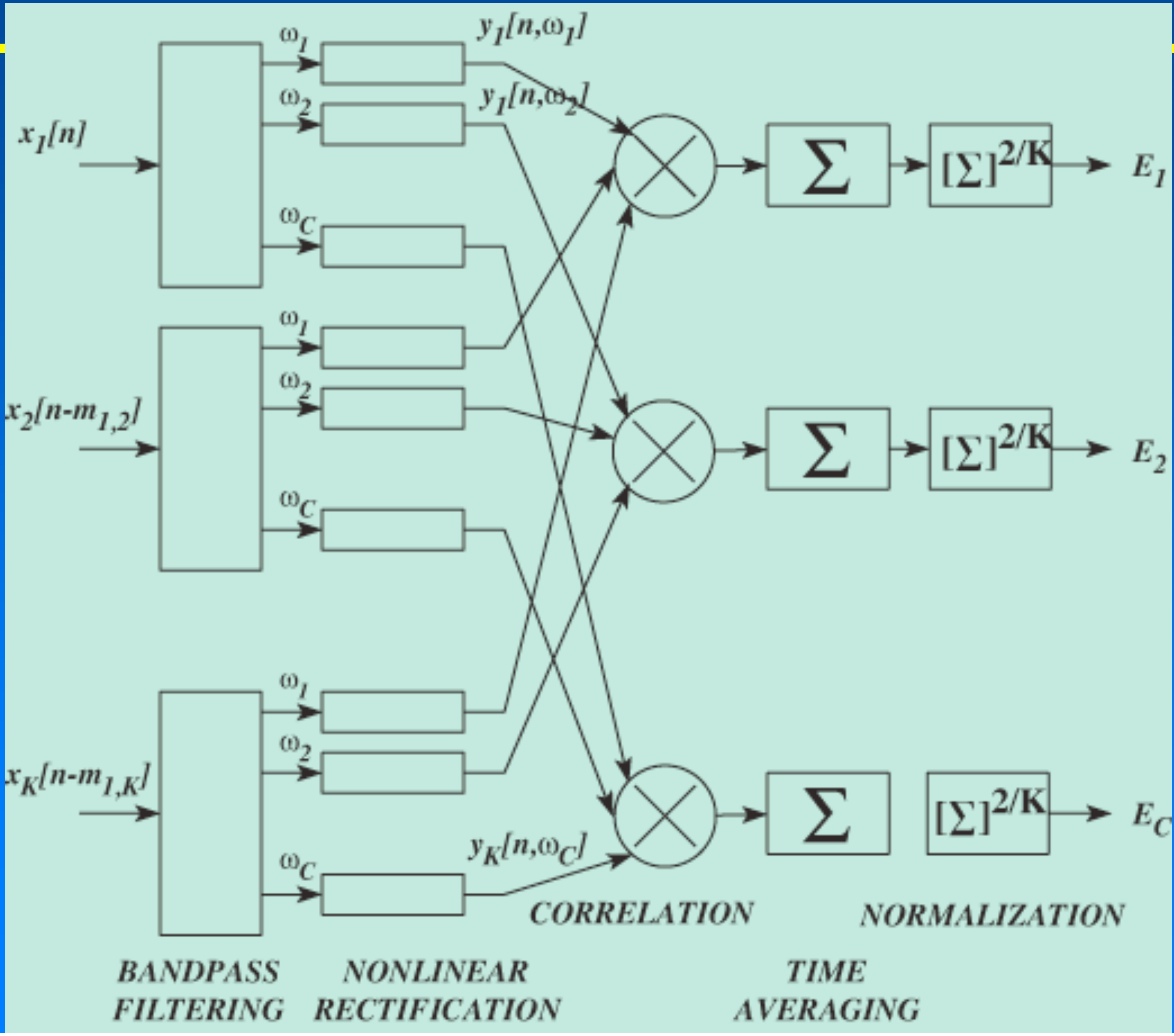
Response to a 500-Hz tone with –1.5-ms ITD



Response to 500-Hz noise with –1.5-ms ITD

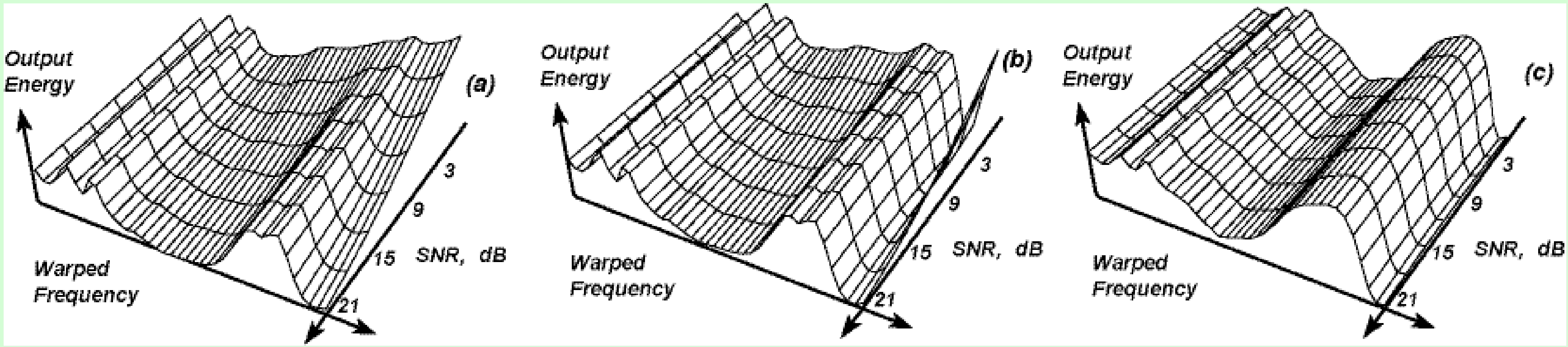


An early application of binaural correlation-based processing to ASR (Sullivan/Stern '93):



The good news: vowel representations improved by correlation processing

- **Reconstructed features of vowel /a/**



Two inputs zero delay | Two inputs 120-ms delay | Eight inputs 120-ms delay

- **Recognition results in 1993 showed some (small) improvement in WER at great computational cost**

So what do things sound like on the cross-correlation display?

- Signals combined with ITDs of 0 and .5 ms
- Individual speech signals:
- Combined speech signals:
- Signals “separated” by correlation display:
- Signals separated by additional correlations across frequency at a common ITD (for “straightness” weighting):

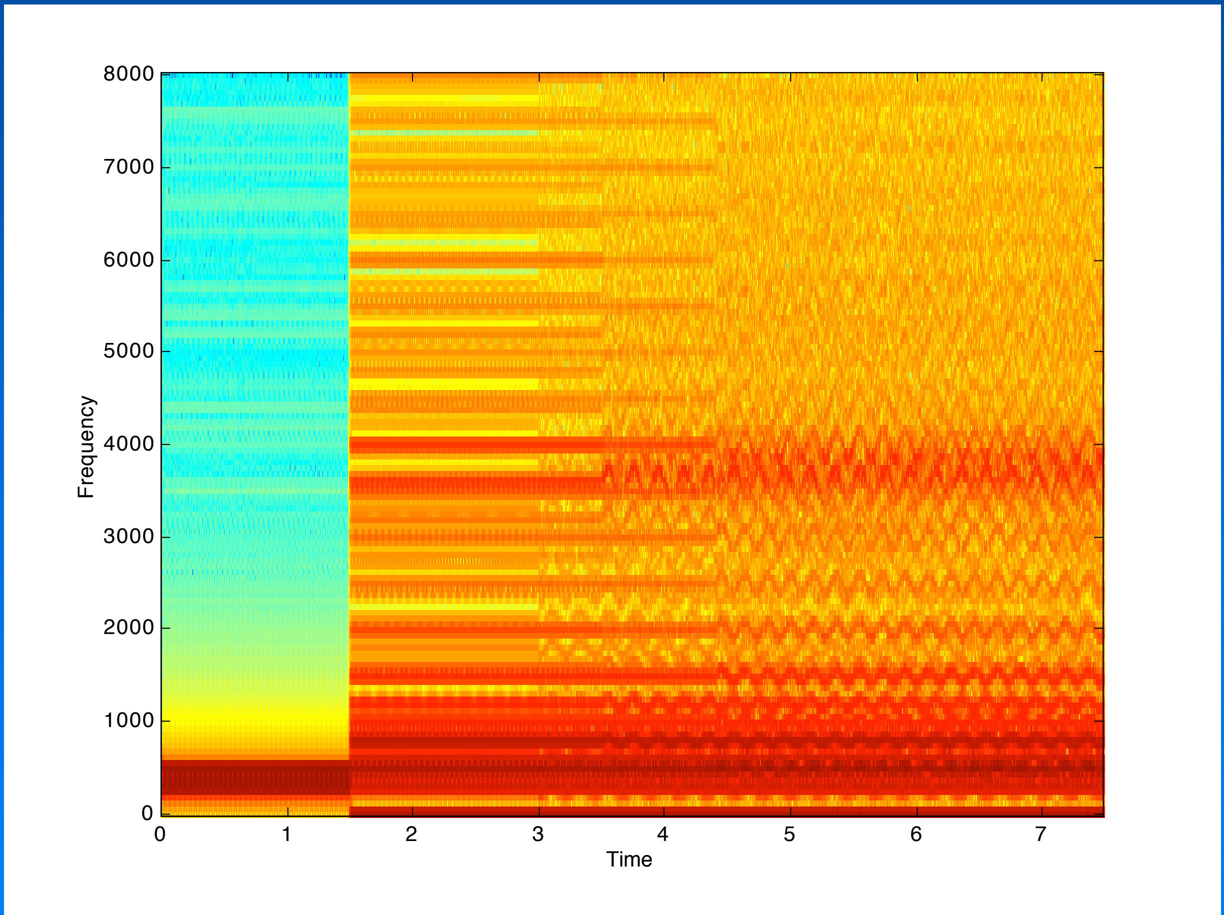
Reverberation remains a difficult problem

- Many modeling efforts are motivated by a desire to account for the precedence effect (*e.g.* Lindemann and others)
- Currently not known whether the precedence effect requires inhibition at the level of the cross-correlation mechanism, or whether it can be accounted for by peripheral auditory-nerve patterns
- Conventional (non-auditory) processing has had some success, but at high computational cost
- Wang and others have used correlation-based processing to isolate spectro-temporal regions that are least likely to be corrupted by reverberation

Signal separation using micro-modulation

- Micromodulation of amplitude and frequency may be helpful in separating unvoiced segments of sound sources
- Physical cues supported by many psychoacoustical studies in recent years

John Chowning's demonstration of effects of micro-modulation in frequency



(Reconstruction based on description in Bregman book)

Separating by frequency modulation only

- **Extract instantaneous frequencies of filterbank outputs**
- **Cross-correlate frequencies across channels (finds co-modulated harmonics)**
- **Cluster correlated harmonics and resynthesize**
- **Our first example:**
  - Isolated speech:
  - Combined speech:
  - Speech separated by frequency modulation:
- **Comment: Success will depend on ability to “track” frequency components across analysis bands**

So why haven't auditory-based representations been more successful to date?

- **Computational complexity** **(at least historically)**
- **Ignoring other information** **besides classical spectral cues**
- **Mismatches** **between extracted features and speech recognition systems**
  – Non-Gaussian probability densities
  – Frame-by-frame temporal analysis

- **A marriage between creative system design and creative signal processing is needed**

Summary and observations

- **Greater computational resources enable us to extract more robust representations based on ongoing timing information**
- **Computational auditory approaches have the potential of providing help in ameliorating some of the most difficult speech recognition problems:**
  - Low SNRs
  - Speech masked by speech and music
  - Reverberant environments
- **But we still need to:**
  - Detect F0 reliably, especially in the presence of competing sources
  - Detect modulations of amplitude and frequency in narrowband channels reliably
  - Track, identify, and disjoint pieces that represent a common source

[redacted]