Robust Automatic Speech Recognition In the 21st Century

Richard Stern

(with Alex Acero, Yu-Hsiang Chiu, Evandro Gouvêa, Chanwoo Kim, Kshitiz Kumar, Amir Moghimi, Pedro Moreno, Hyung-Min Park, Bhiksha Raj, Mike Seltzer, Rita Singh, and many others)

Robust Speech Recognition Group
Carnegie Mellon University

Telephone: +1 412 268-2535
Fax: +1 412 268-3890
rms@cs.cmu.edu
http://www.ece.cmu.edu/~rms

AFEKA Conference for Speech Processing
Tel Aviv, Israel
July 7, 2014

Robust speech recognition

- **As speech recognition is transferred from the laboratory to the marketplace robust recognition is becoming increasingly important**
- **"Robustness" in 1985:**
  - Recognition in a quiet room using desktop microphones
- **Robustness in 2014:**
  - Recognition ….
    - » over a cell phone
    - » in a car
    - » with the windows down
    - » and the radio playing
    - » at highway speeds

What I would like to do today

- **Review background and motivation for current work:**
  - Sources of environmental degradation
- **Discuss selected approaches and their performance:**
  - Traditional statistical parameter estimation
  - Missing feature approaches
  - Microphone arrays
  - Physiologically- and perceptually-motivated signal processing
- **Comment on current progress for the hardest problems**

Some of the hardest problems in speech recognition

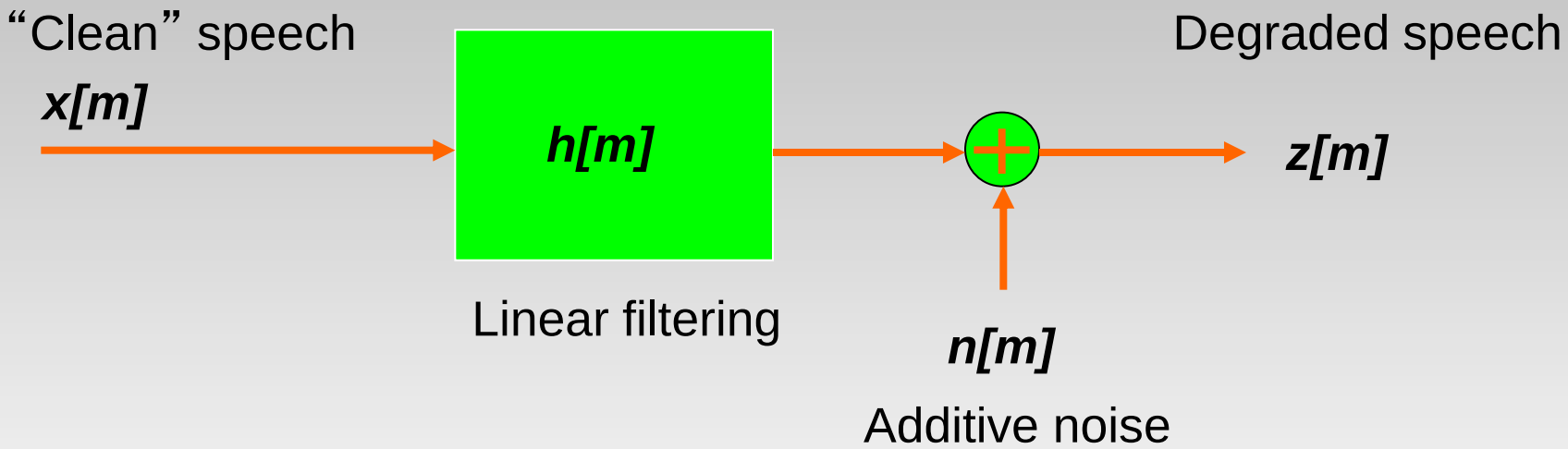
- Speech in high noise (Navy F-18 flight line)
- Speech in background music
- Speech in background speech
- Transient dropouts and noise
- Spontaneous speech
- Reverberated speech
- Vocoded speech

Challenges in robust recognition

- **"Classical" problems:**
  - Additive noise
  - Linear filtering
- **"Modern" problems:**
  - Transient degradations
  - Much lower SNR
- **"Difficult" problems:**
  - Highly spontaneous speech
  - Reverberated speech
  - Speech masked by other speech and/or music
  - Speech subjected to nonlinear degradation

Solutions to classical problems: joint statistical compensation for noise and filtering

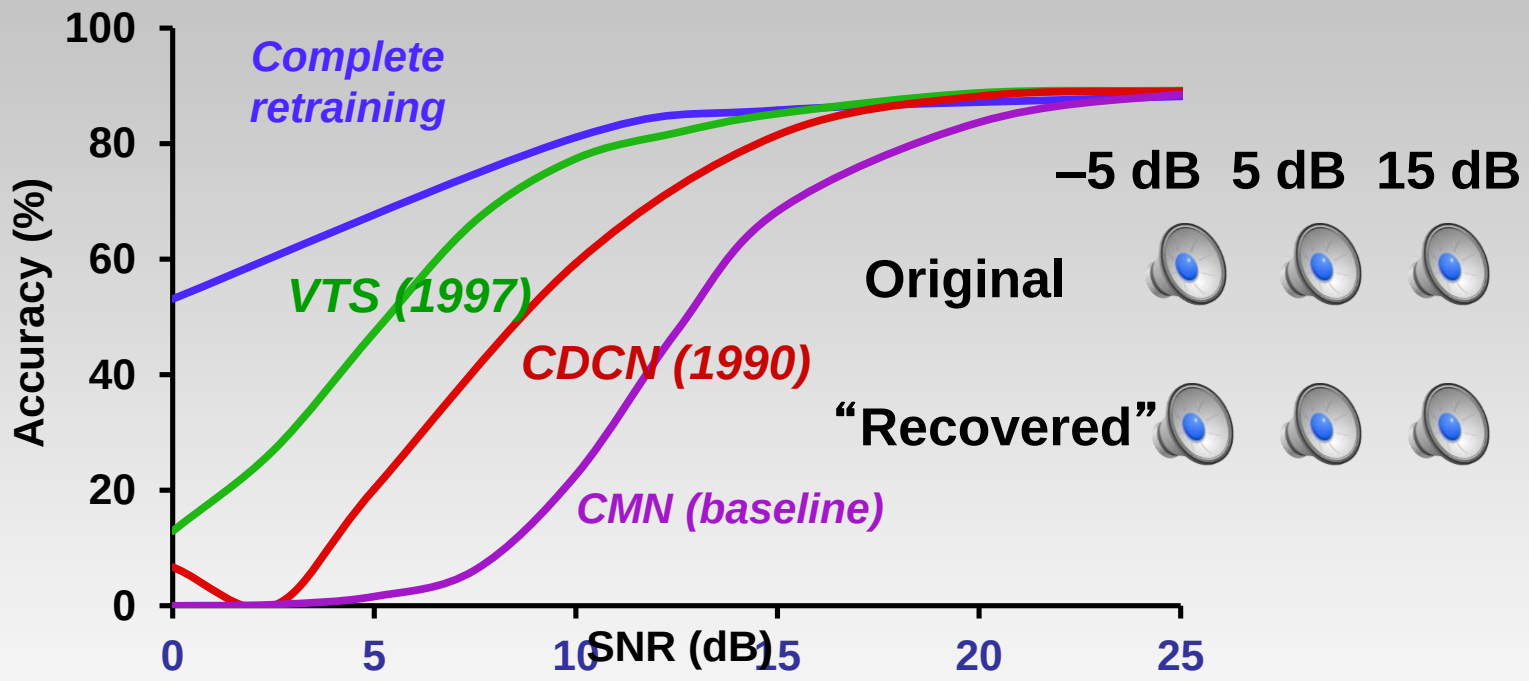
- **Approach of Acero, Liu, Moreno, and Raj, et al. (1990-1997)…**



- **Compensation achieved by estimating parameters of noise and filter and applying inverse operations**
- **Interaction is nonlinear:**

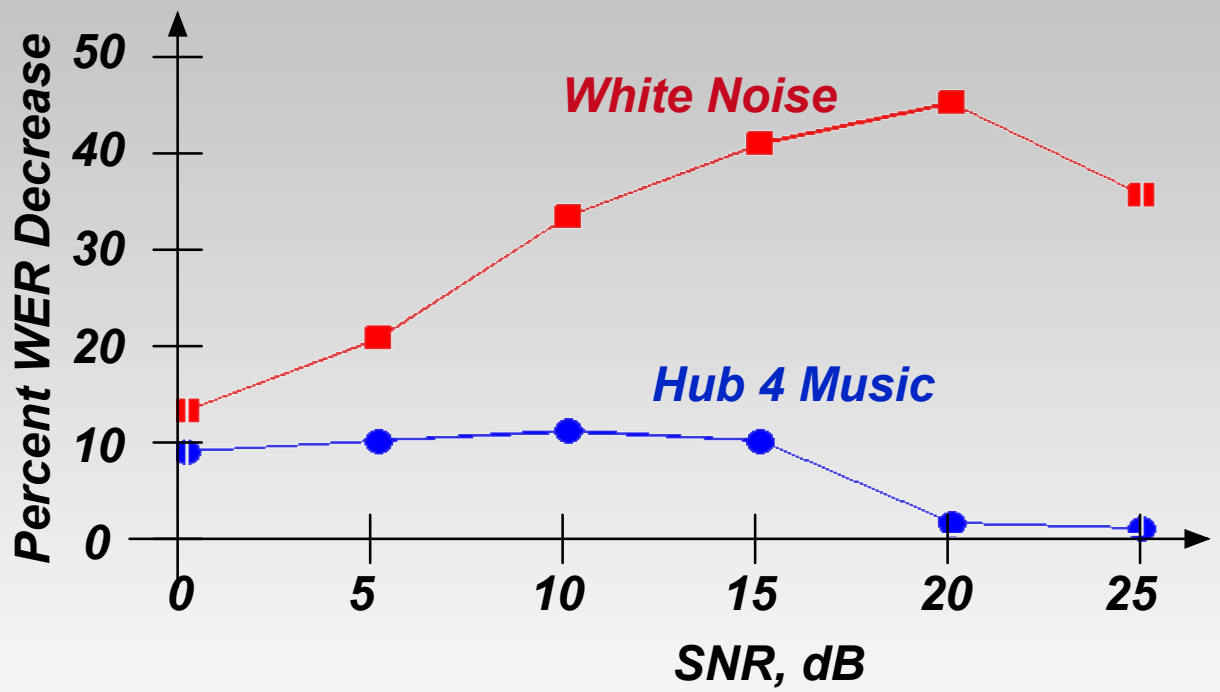
$$z = x + q + \log\left(1 + e^{n-x-q}\right)$$

“Classical” combined 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

Summary: traditional methods

- The effects of additive noise and linear filtering are nonlinear
- Methods such as CDCN and VTS can be quite effective, but …
- … these methods require that the statistics of the received signal remain stationary over an interval of several hundred ms

Introduction: Missing-feature recognition

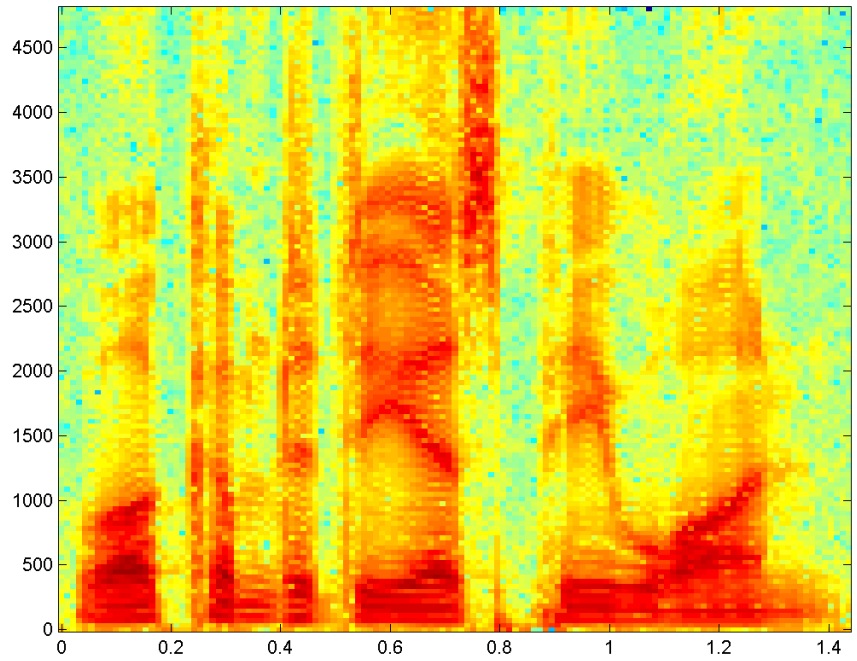
- **Speech is quite intelligible, even when presented only in fragments**
- **Procedure:**
  - Determine which time-frequency time-frequency components appear to be unaffected by noise, distortion, etc.
  - Reconstruct signal based on “good” components
- **A monaural example using “oracle” knowledge:**
  - Mixed signals -
  - Separated signals -

Missing-feature recognition

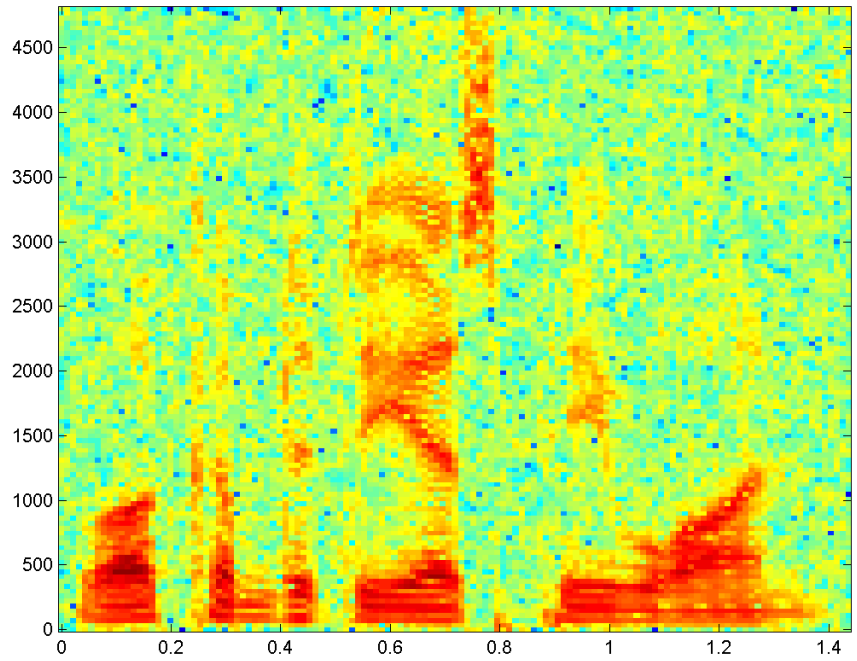
- **General approach:**
  - Determine which cells of a spectrogram-like display are unreliable (or "missing")
  - Ignore missing features or make best guess about their values based on data that are present

- **Comment:**
  - Most groups (following the University of Sheffield) modify the internal representations to compensate for missing features
  - We attempt to infer and replace missing components of input vector

Example: an original speech spectrogram

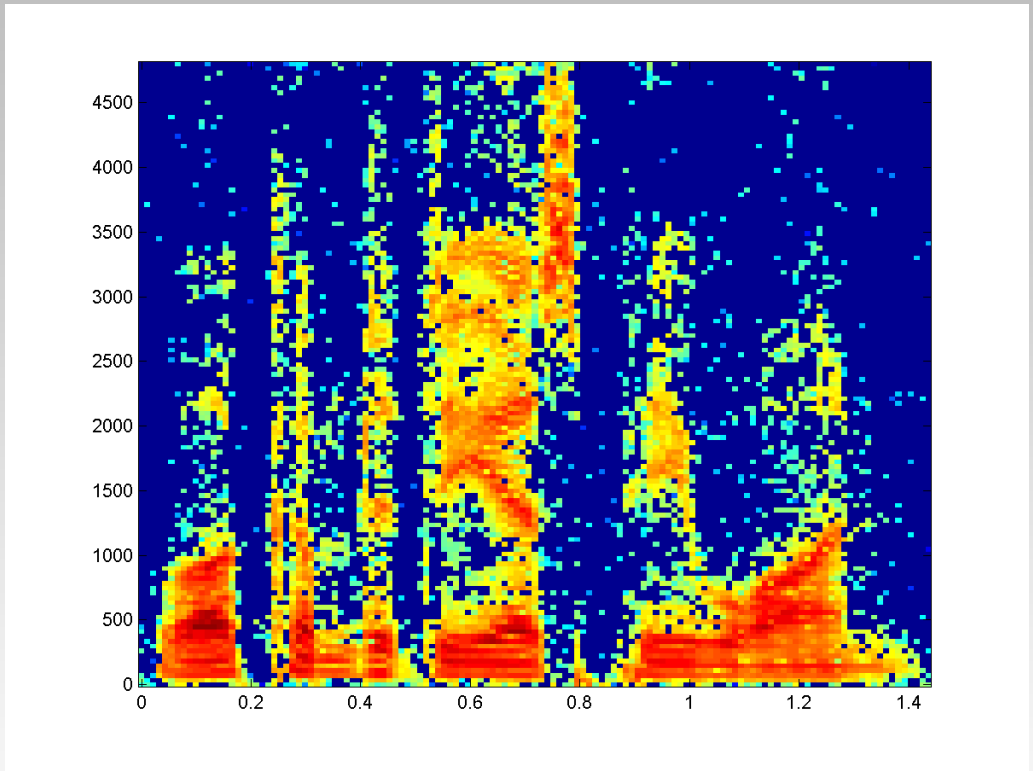


Spectrogram corrupted by noise at SNR 15 dB



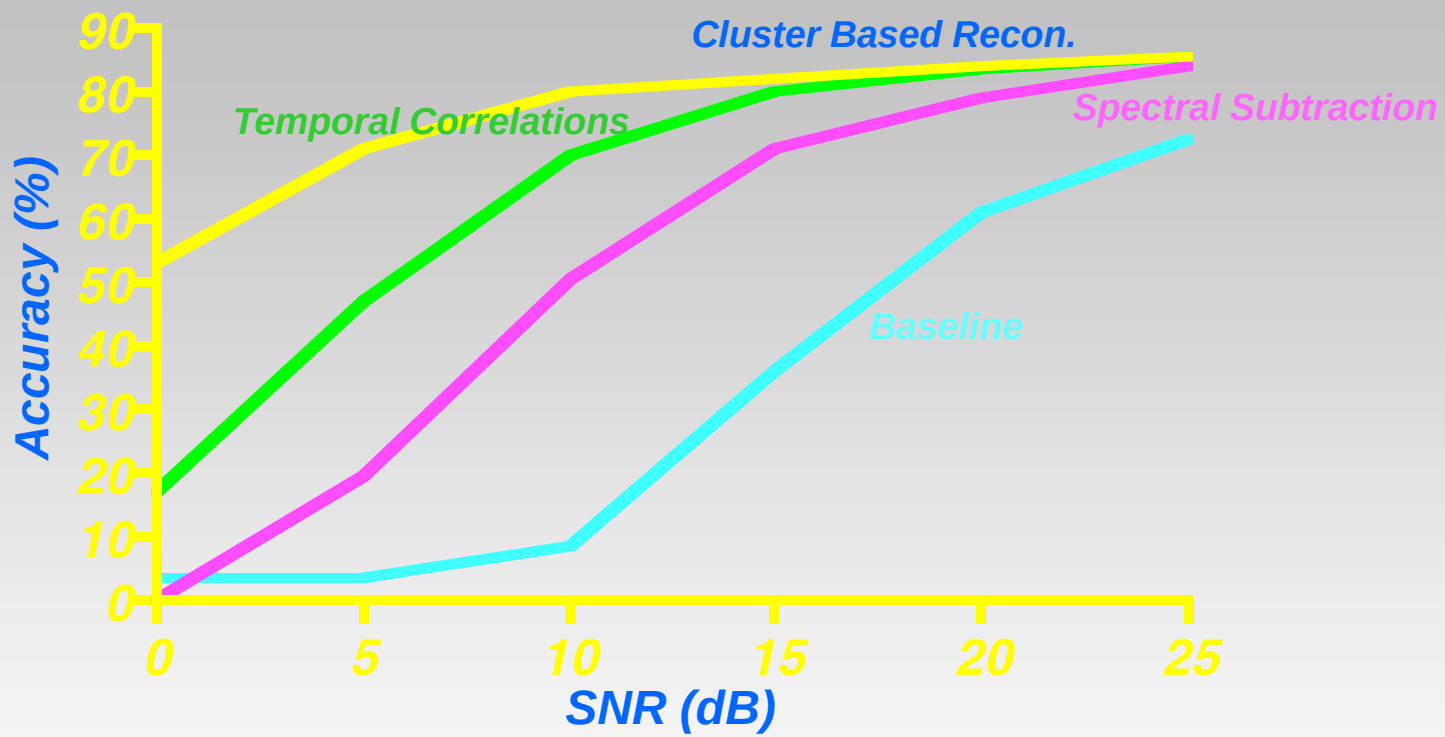
Some regions are affected far more than others

Ignoring regions in the spectrogram that are corrupted by noise



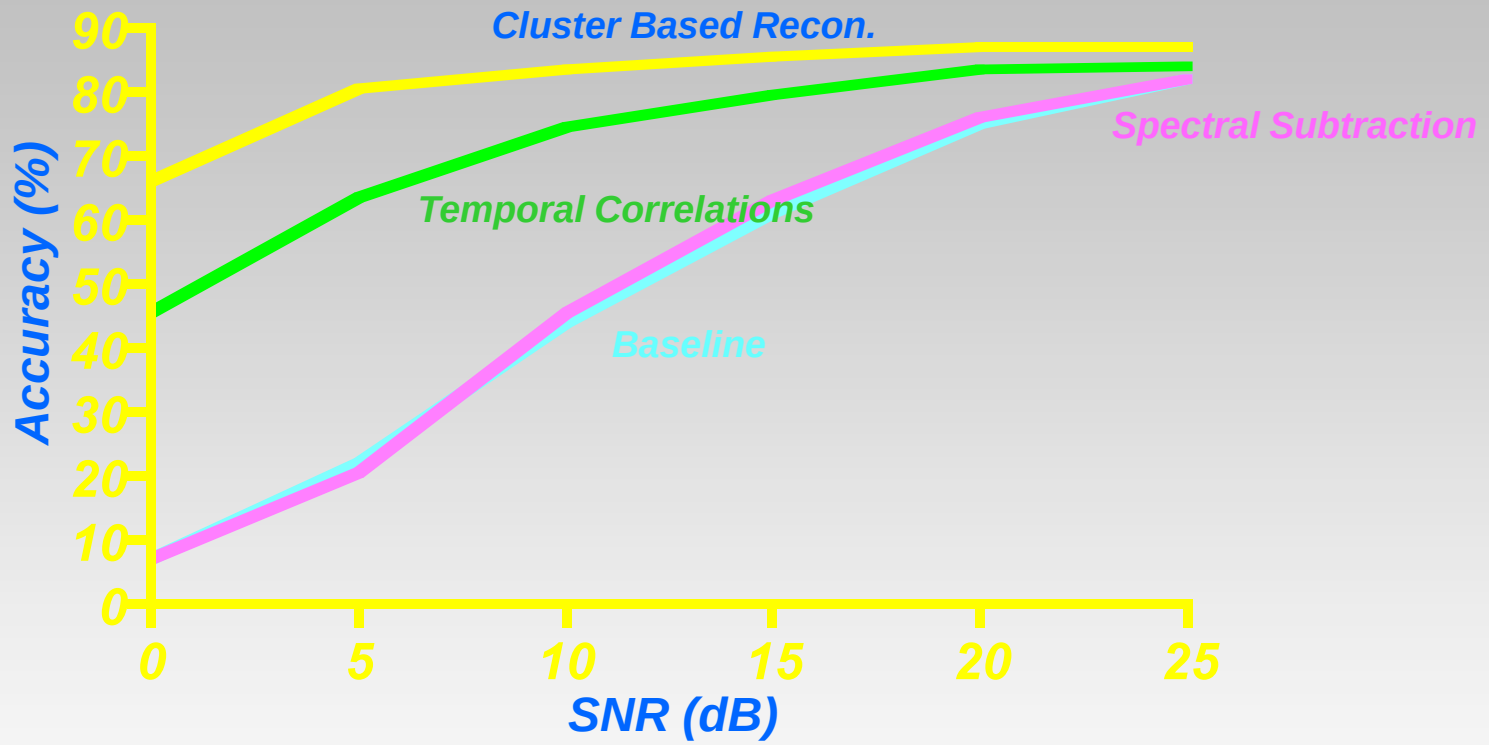
- All regions with SNR less than 0 dB deemed missing (dark blue)
- Recognition performed based on colored regions alone

Recognition accuracy using compensated cepstra, speech in white noise (Raj, 1998)



- Large improvements in recognition accuracy can be obtained by reconstruction of corrupted regions of noisy speech spectrograms
- *A priori k*nowledge of locations of "missing" features needed

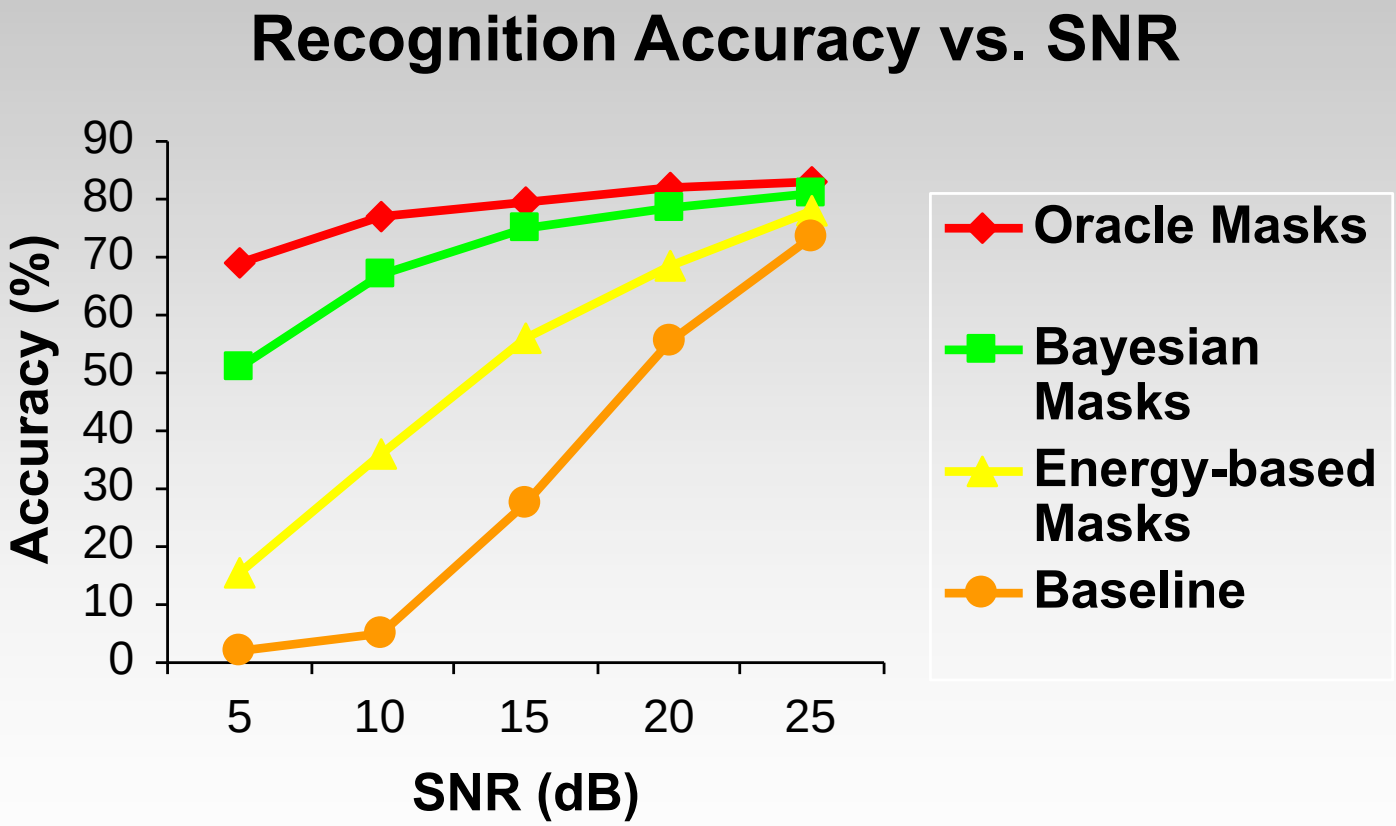
Recognition accuracy using compensated cepstra, speech corrupted by music



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

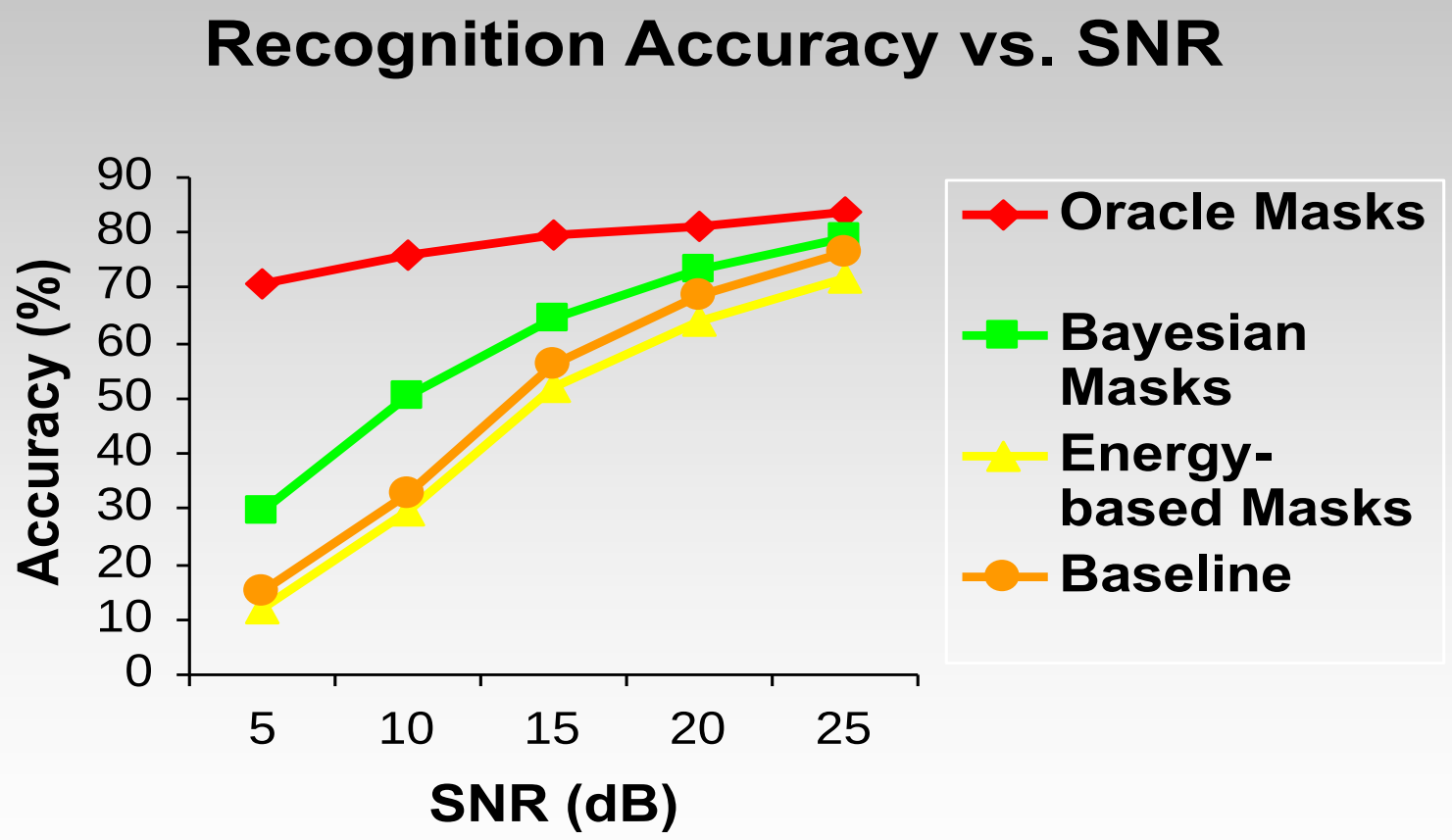
Practical recognition error: white noise (Seltzer, 2000)

- **Speech plus white noise:**



Practical recognition error: background music

- **Speech plus music:**



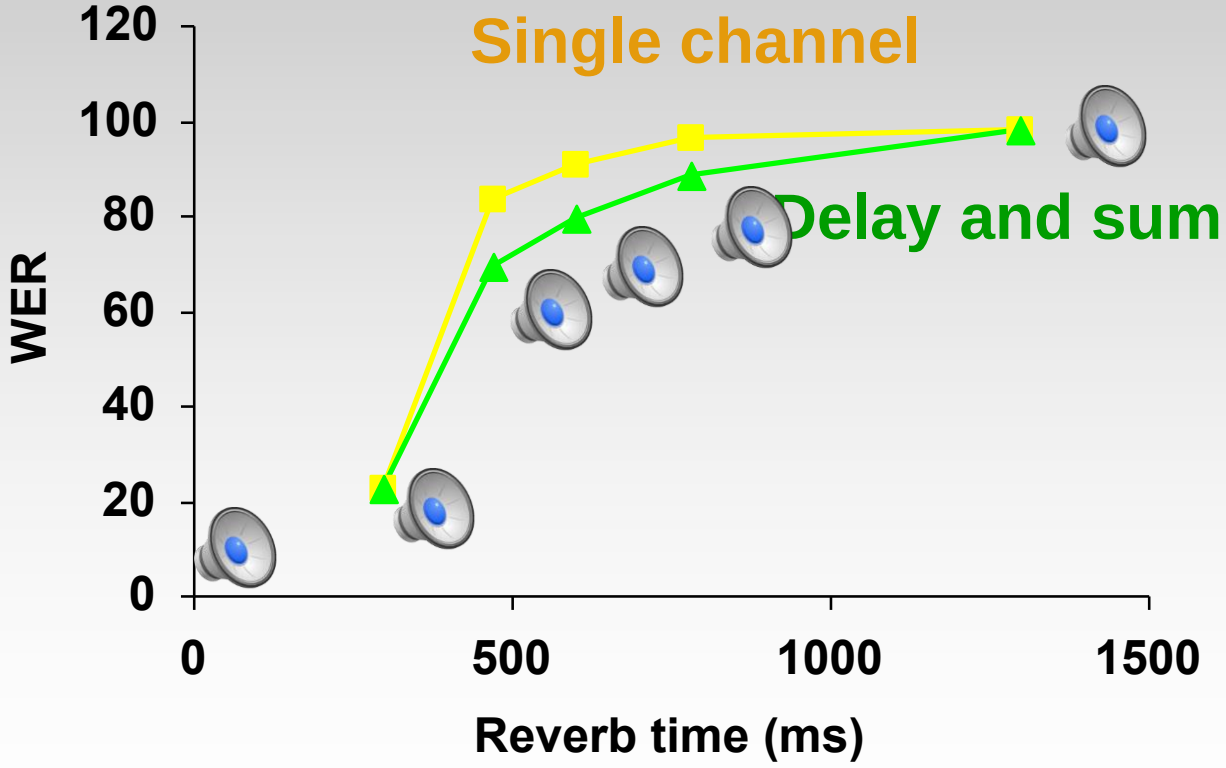
Summary: Missing features

- **Missing feature approaches can be valuable in dealing with the effects of transient distortion and other disruptions that are localized in the spectro-temporal display**

- **The approach can be effective, but it is limited by the need to determine correctly which cells in the spectrogram are missing, which can be difficult in practice**

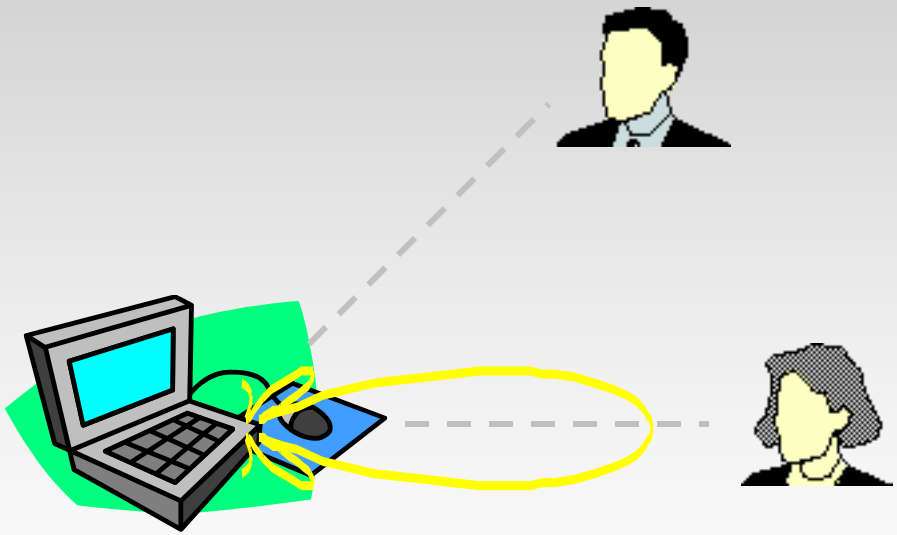
The problem of reverberation

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



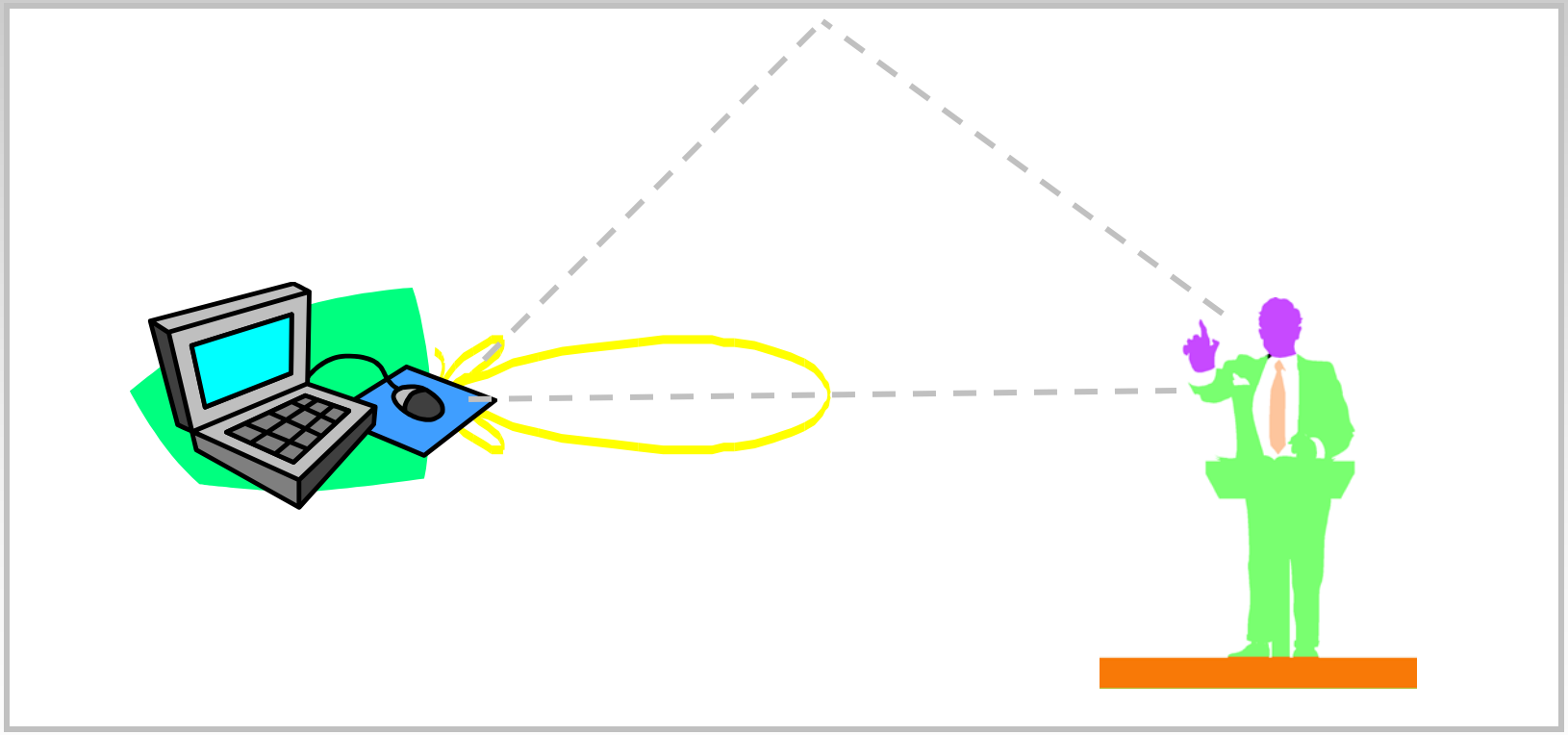
Use of microphone arrays: motivation

- **Microphone arrays can provide directional response, accepting speech from some directions but suppressing others**



Another reason for microphone arrays …

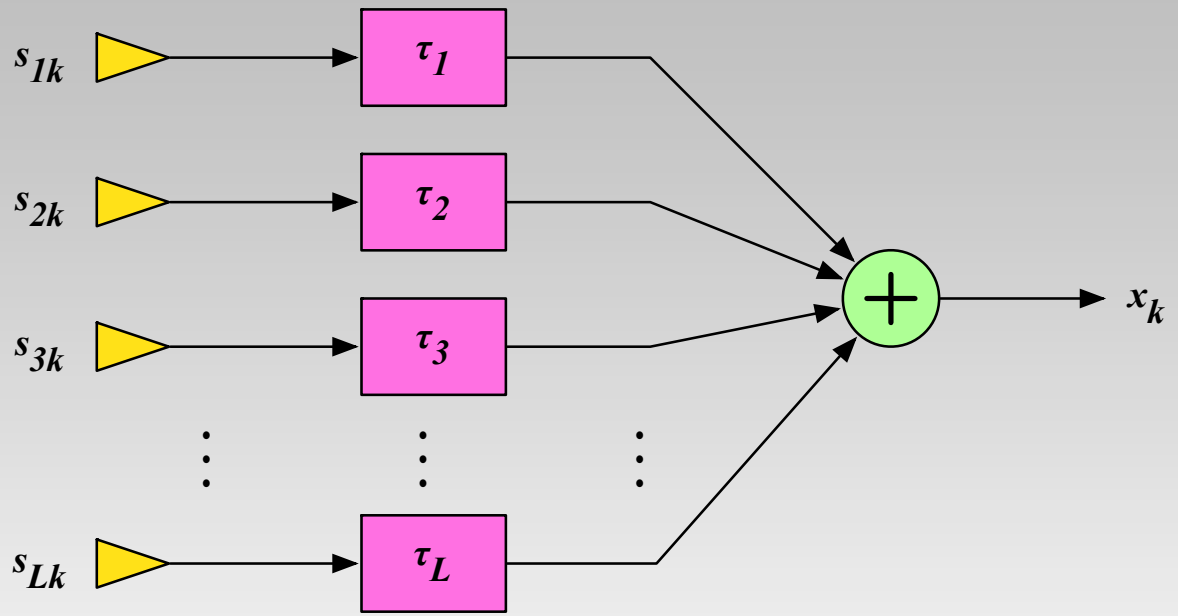
- **Microphone arrays can focus attention on the direct field in a reverberant environment**



Options in the use of multiple microphones…

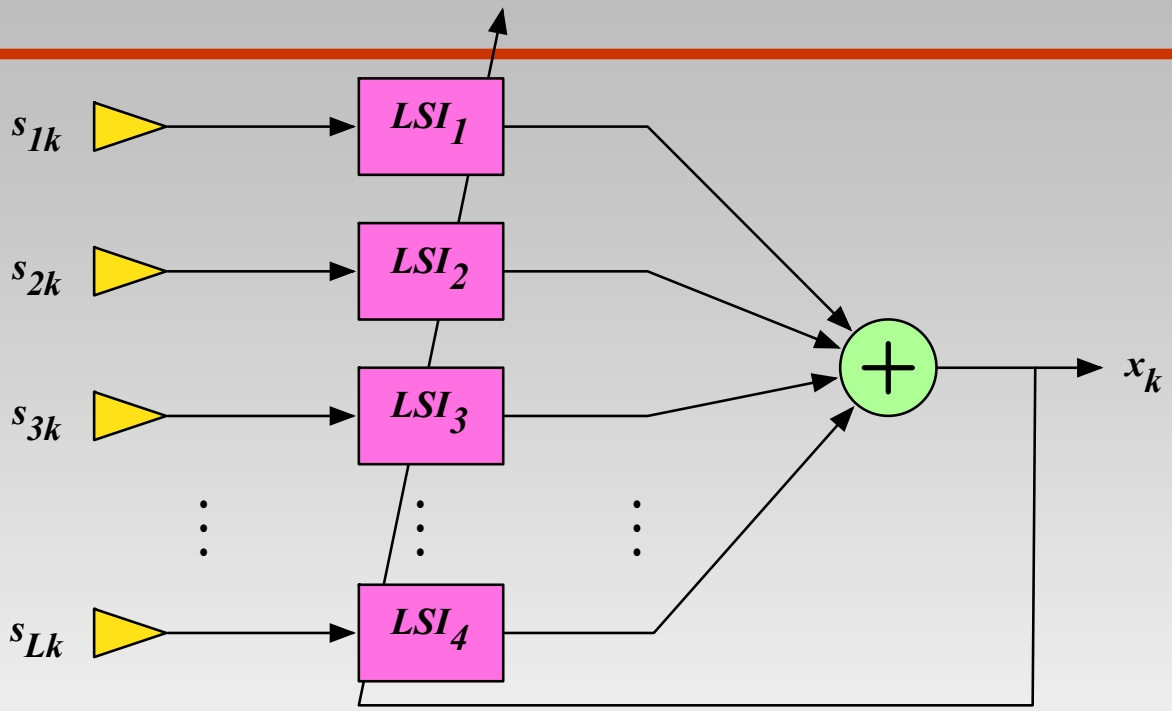
- **There are many ways we can use multiple microphones to improve recognition accuracy:**
  - Fixed delay-and-sum beamforming
  - Microphone selection techniques
  - Traditional adaptive filtering based on minimizing waveform distortion
  - Feature-driven adaptive filtering (LIMABEAM)
  - Statistically-driven separation approaches (ICA/BSS)
  - Binaural processing based on selective reconstruction (*e.g.* PDCW)
  - Binaural processing for correlation-based emphasis (*e.g.* Polyaural)
  - Binaural processing using precedence-based emphasis (peripheral or central, *e.g.* SSF)

Delay-and-sum beamforming



- Simple processing based on equalizing delays to sensors and summing responses
- High directivity can be achieved with many sensors
- Baseline algorithm for any multi-microphone experiment

Adaptive array processing



- **MMSE-based methods (e.g. LMS, RLS) falsely assume independence of signal and noise; not true in reverberation**
  - Not as much of an issue with modern methods using objective functions based on kurtosis or negative entropy
- **Methods reduce signal distortion, not error rate**

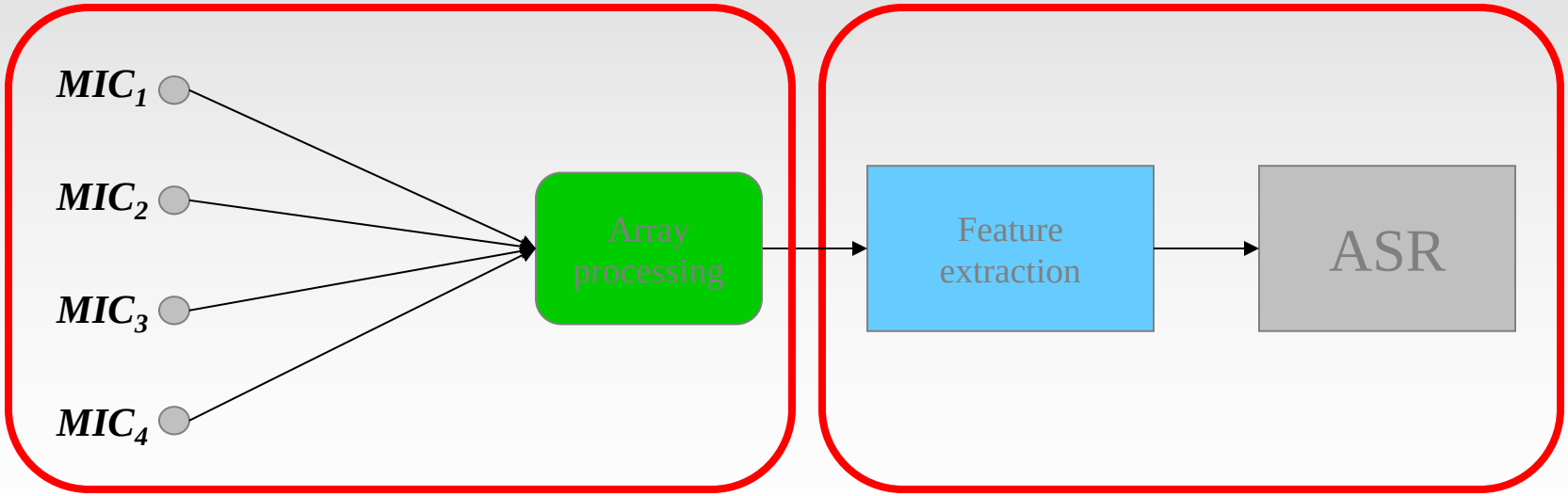
Speech recognition using microphone arrays

- **Speech recognition using microphone arrays has been always been performed by combining two independent systems.**

  This is not ideal:

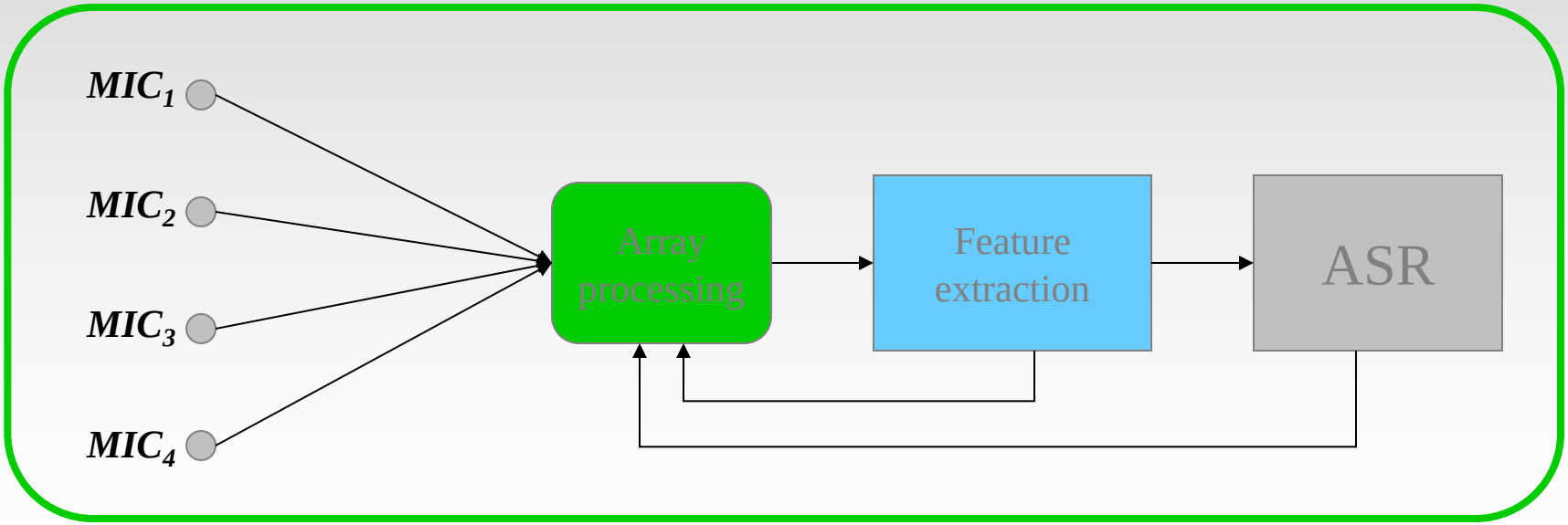
  – Systems have different objectives

  – Each system does not exploit information available to the other

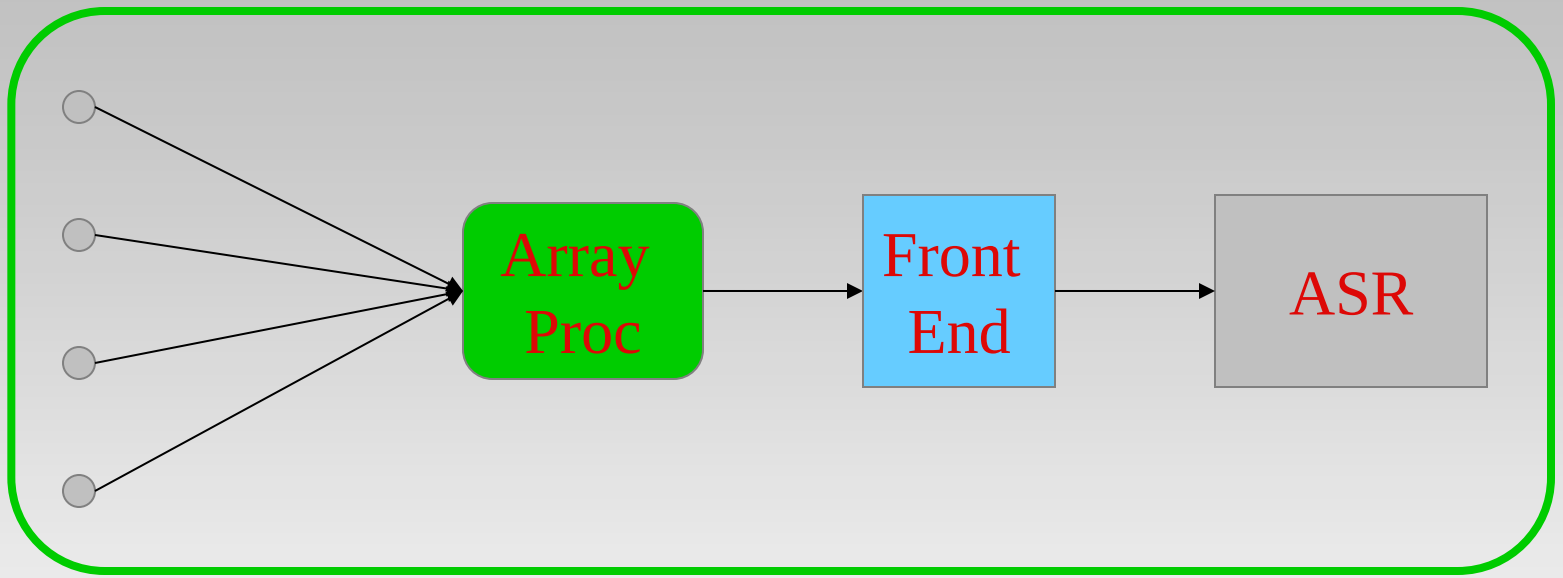


Feature-based optimal filtering (Seltzer 2004)

- Consider array processing and speech recognition as part of a single system that shares information
- Develop array processing algorithms specifically designed to improve speech recognition



Multi-microphone compensation for speech recognition based on cepstral distortion



- **Multi-mic compensation based on optimizing speech features rather than signal distortion**

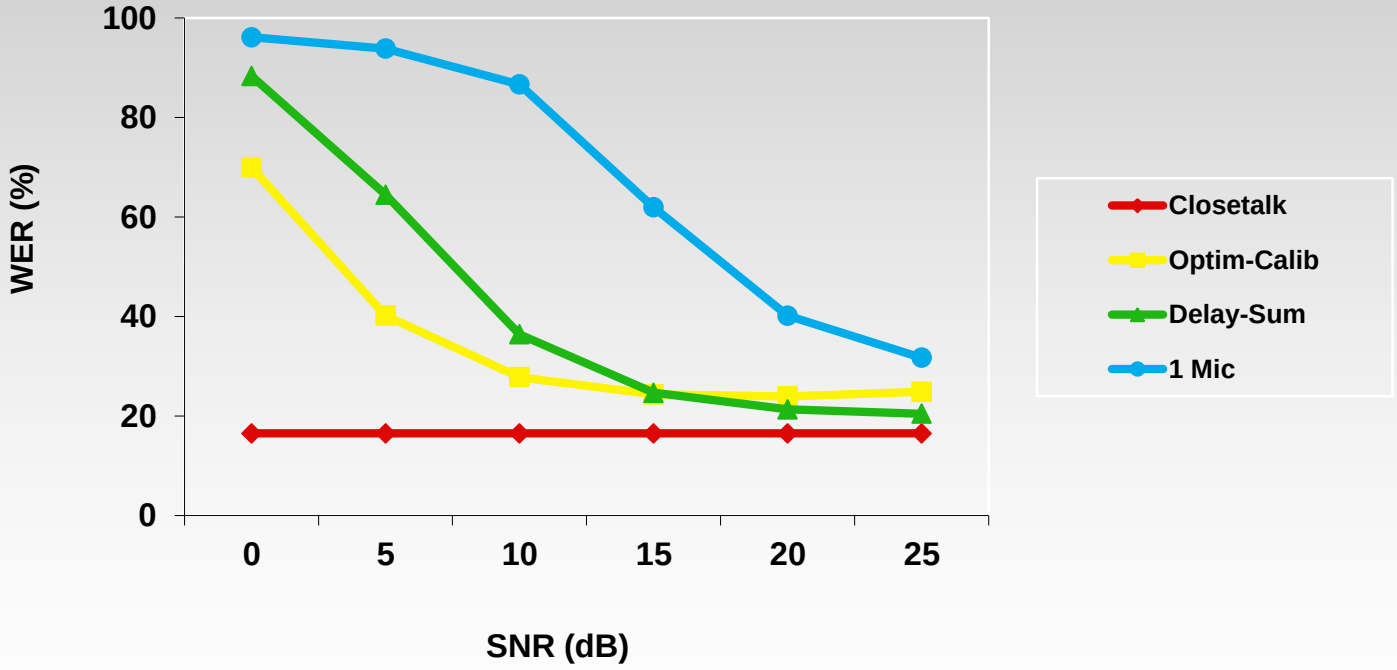
Speech in Room

Delay and Sum

Optimal Comp

Sample results

- **WER vs. SNR for WSJ with added white noise:**
  - Constructed 50-point filters from calibration utterance using transcription only
  - Applied filters to all utterances



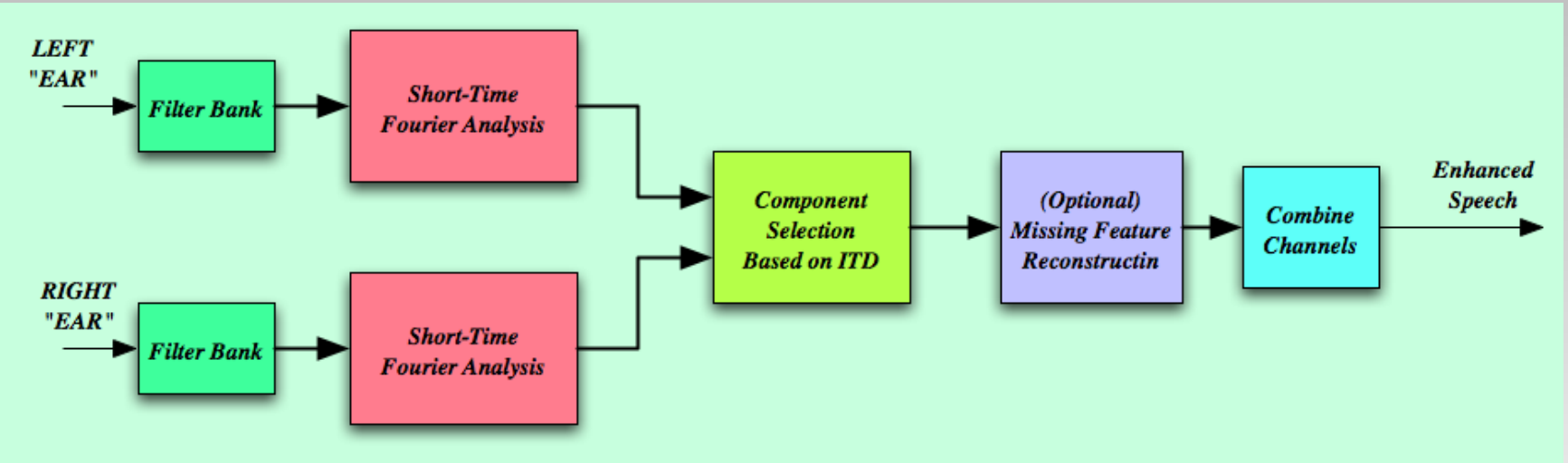
“Nonlinear beamforming:” reconstructing sound from fragments

- **Procedure:**
  - Determine which time-frequency time-frequency components appear to be dominated by the desired signal
  - Recognize based on subset of features that are “good”

  OR

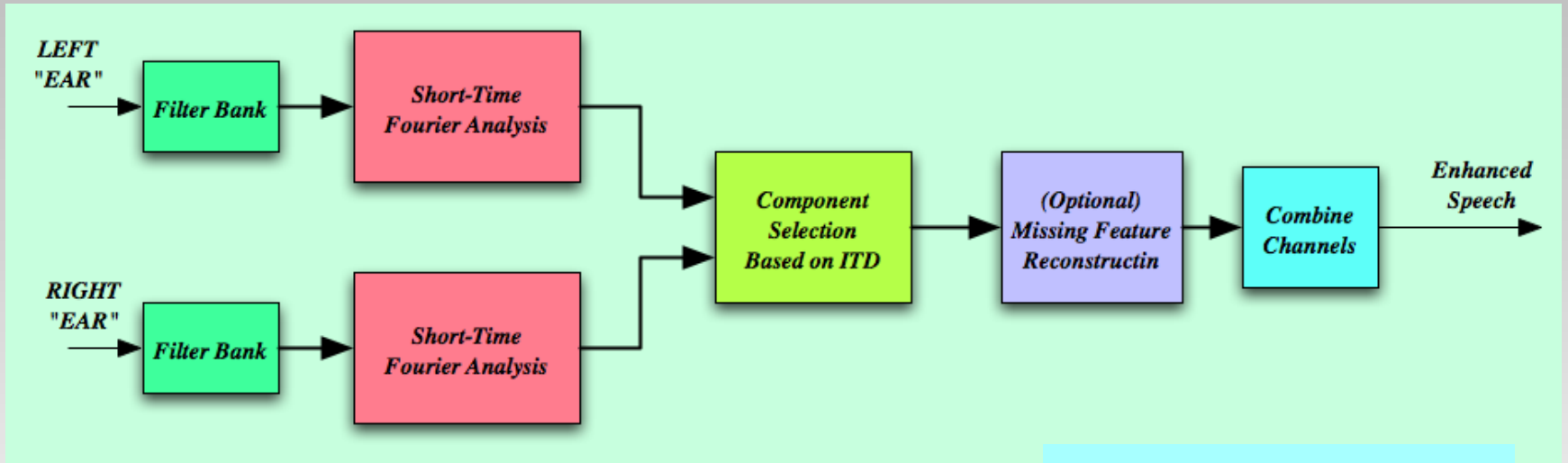
  - Reconstruct signal based on “good” components and recognize using traditional signal processing
  - In binaural processing determination of “good” components is based on estimated ITD

Binaural processing for selective reconstruction (*e.g.* ZCAE, PDCW processing)



- Assume two sources with known azimuths
- Extract ITDs in TF rep (using zero crossings, cross-correlation, or phase differences in frequency domain)
- Estimate signal amplitudes based on observed ITD (in binary or continuous fashion)
- (Optionally) fill in missing TF segments after binary decisions

Audio samples using selective reconstruction

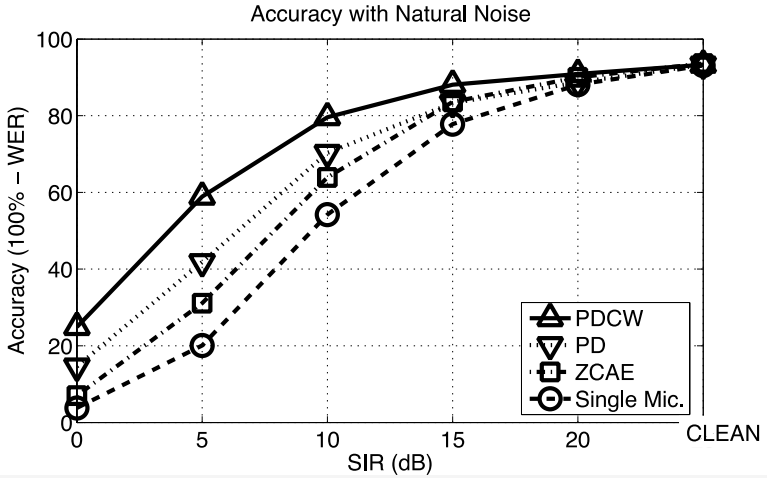


RT60 (ms)	300
No Proc	
Delay-sum	
PDCW	

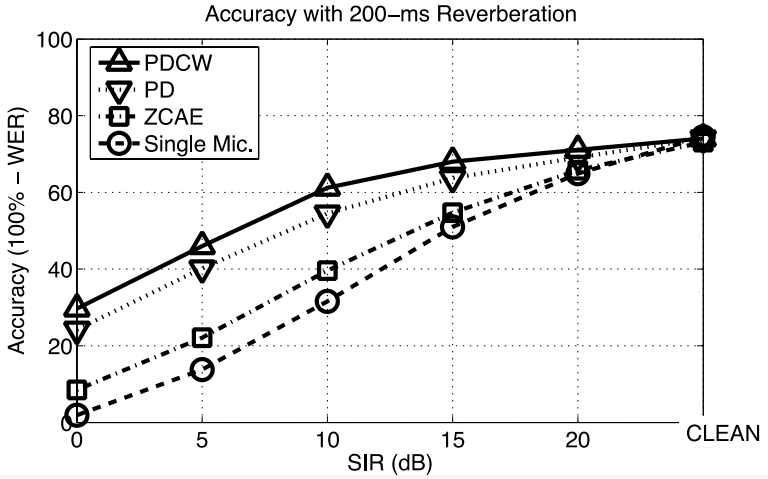
Selective reconstruction from two mics helps

- **Examples using the PDCW algorithm (Kim et al. Interspeech 2009):**

Speech in “natural” noise:



Reverberated speech:

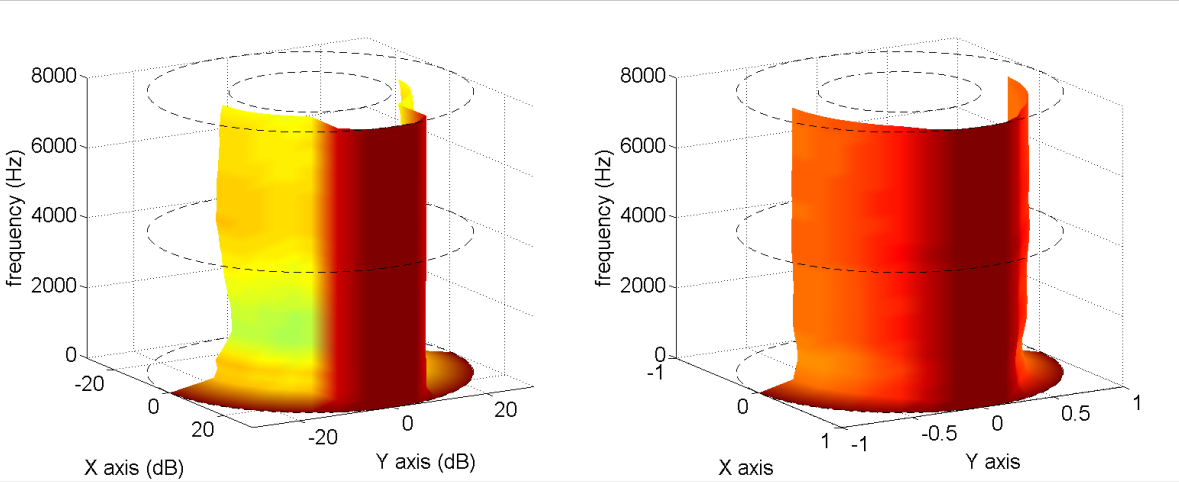


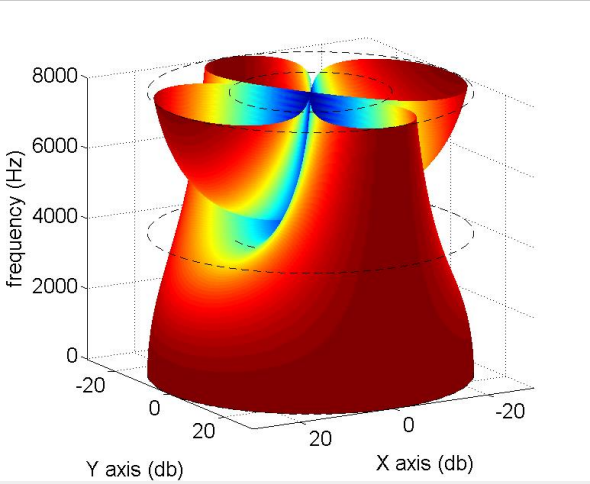
- **Comment:** Use of two mics provides substantial improvement that is typically independent of what is obtained using other methods

Comparing linear and "nonlinear" beamforming

- **"Nonlinear beampatterns:**

Linear beampattern:





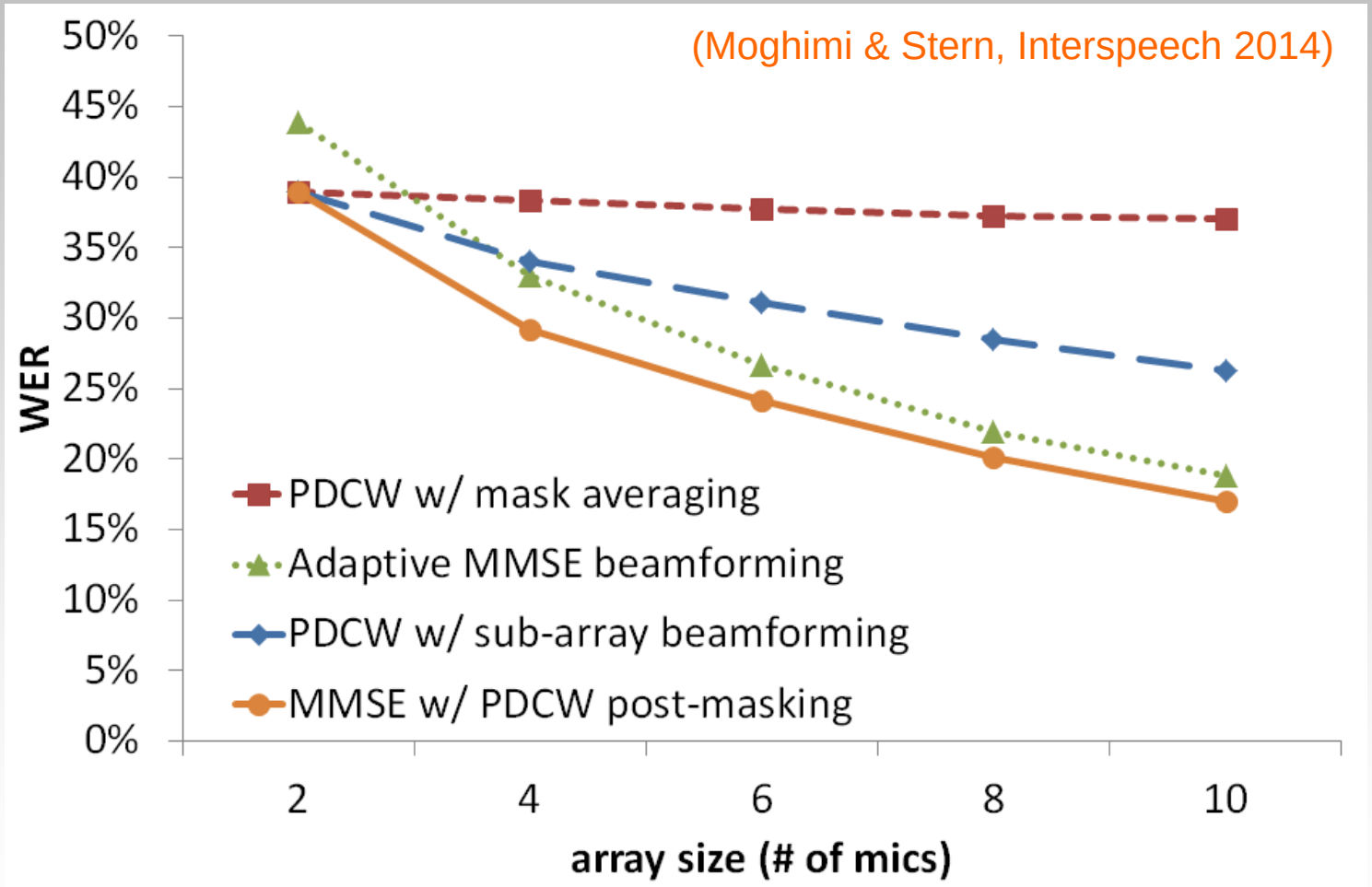
SNR = 0 dB ***SNR = 20 dB***

- **Comments:**

(Moghimi, ICASSP 2014)

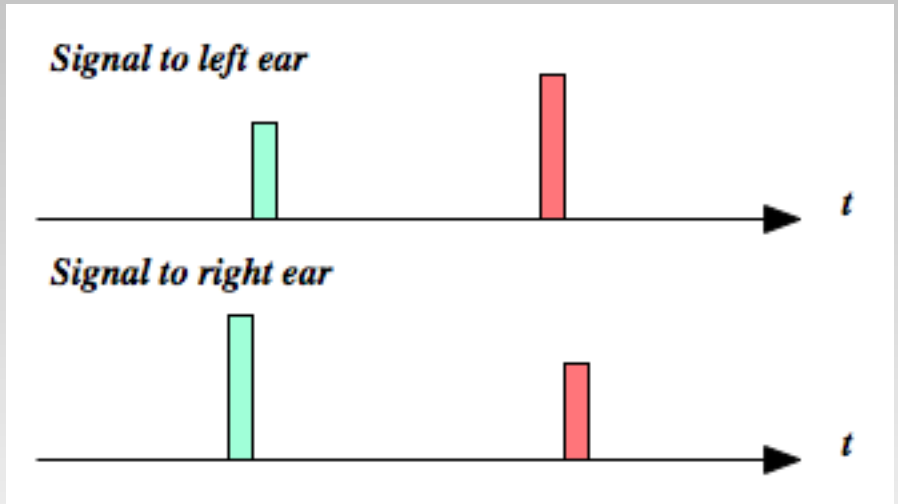
  - Performance depends on SNR as well as source locations
  - More consistent over frequency than linear beamforming

Linear and nonlinear beamforming as a function of the number of sensors



The binaural precedence effect

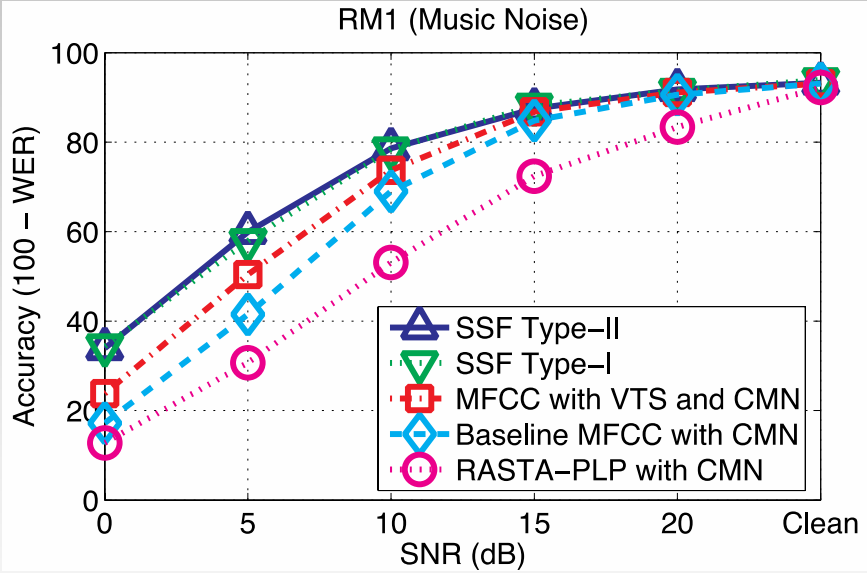
- Basic stimuli of the precedence effect:



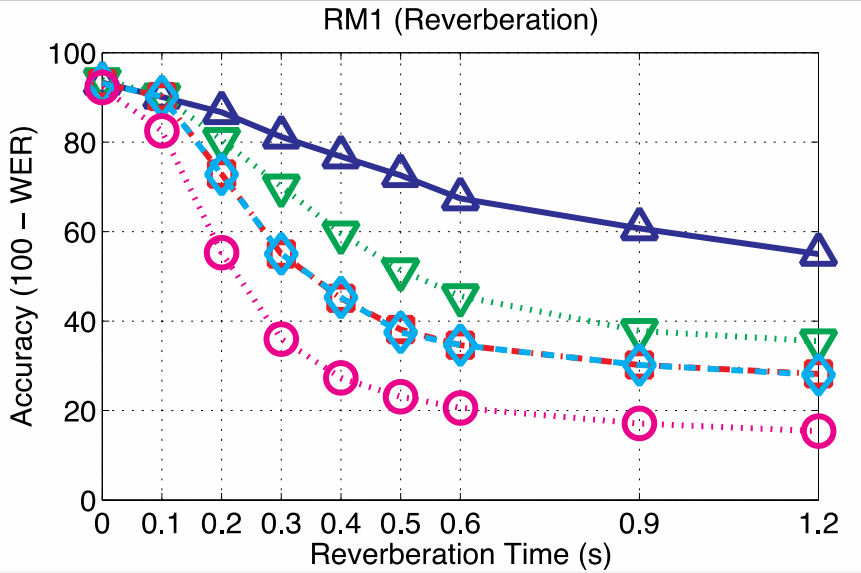
- Localization is typically dominated by the first arriving pair
- Precedence effect believed by some (e.g. Blauert) to improve speech intelligibility
- Generalizing, we might say that onset enhancement helps at any level

Performance of onset enhancement in the SSF algorithm (Kim and Stern, Interspeech 2010)

- **Background music:**

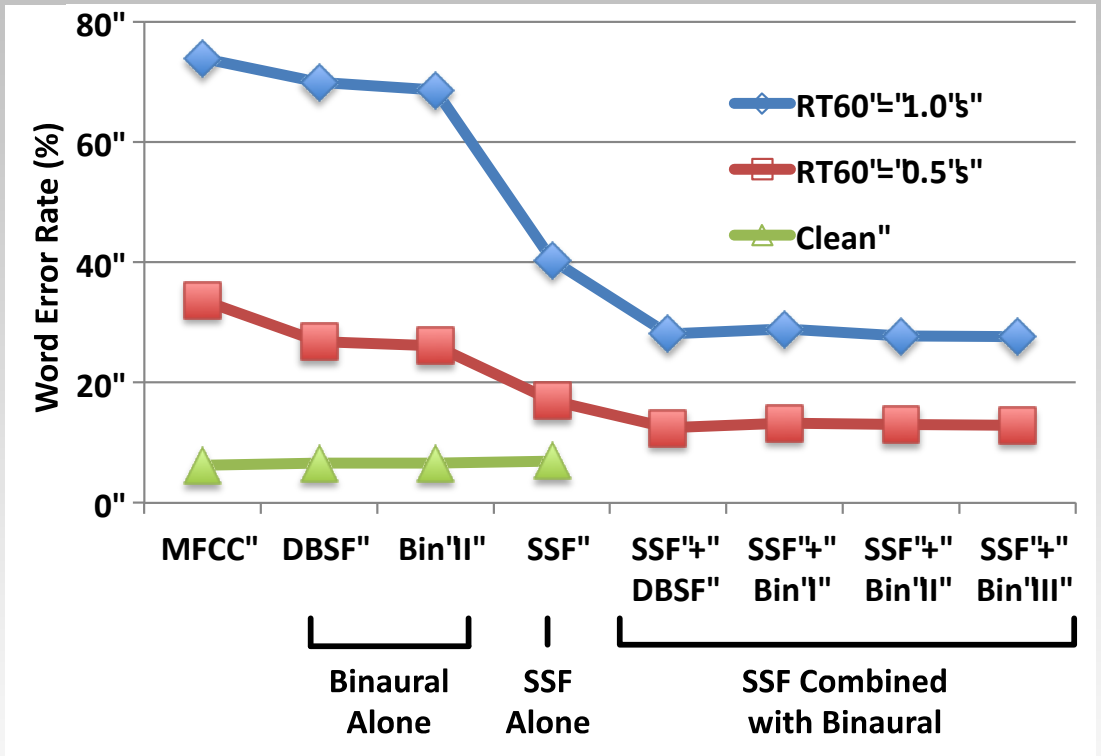


- **Reverberation:**



Comment: Onset enhancment using SSF processing is especially effective in dealing with reverberation

Combining onset enhancement with two-microphone processing:



(Park et al., Interspeech 2014)

- **Comment:** the use of both SSF onset enhancement and binaural comparison is especially helpful for improving WER for reverberated speech

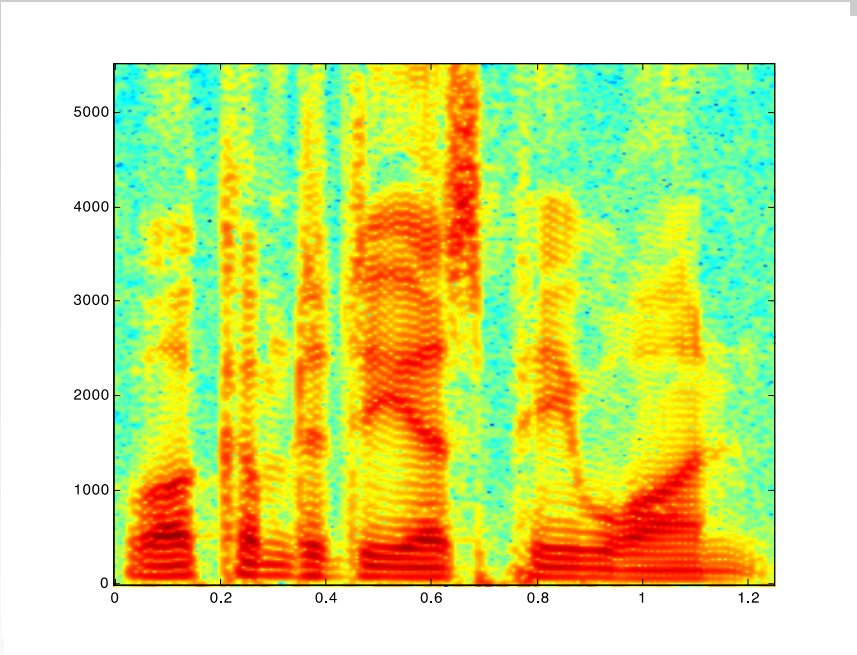
Summary: use of multiple mics

- Microphone arrays provide directional response which can help interfering sources and in reverberation
- Delay-and-sum beamforming is very simple and somewhat effective
- Adaptive beamforming provides better performance but not in reverberant environments with MMSE-based objective functions
- Adaptive beamforming based on minimizing feature distortion can be very effective but is computationally costly
- For only two mics, “nonlinear” beamforming based on selective reconstruction is best
- Onset enhancement helps a great deal as well in reverberance

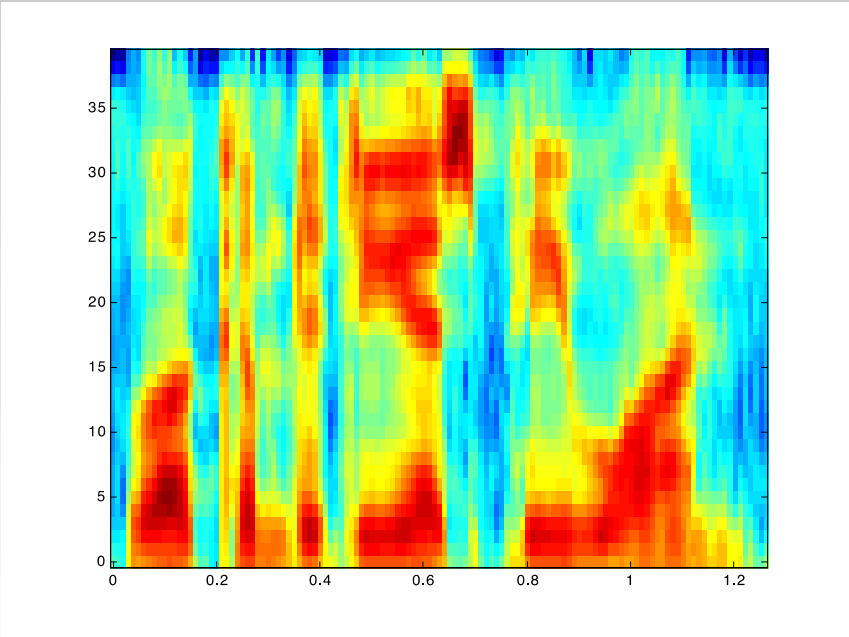
Auditory-based representations

What the speech recognizer sees:

An original spectrogram:



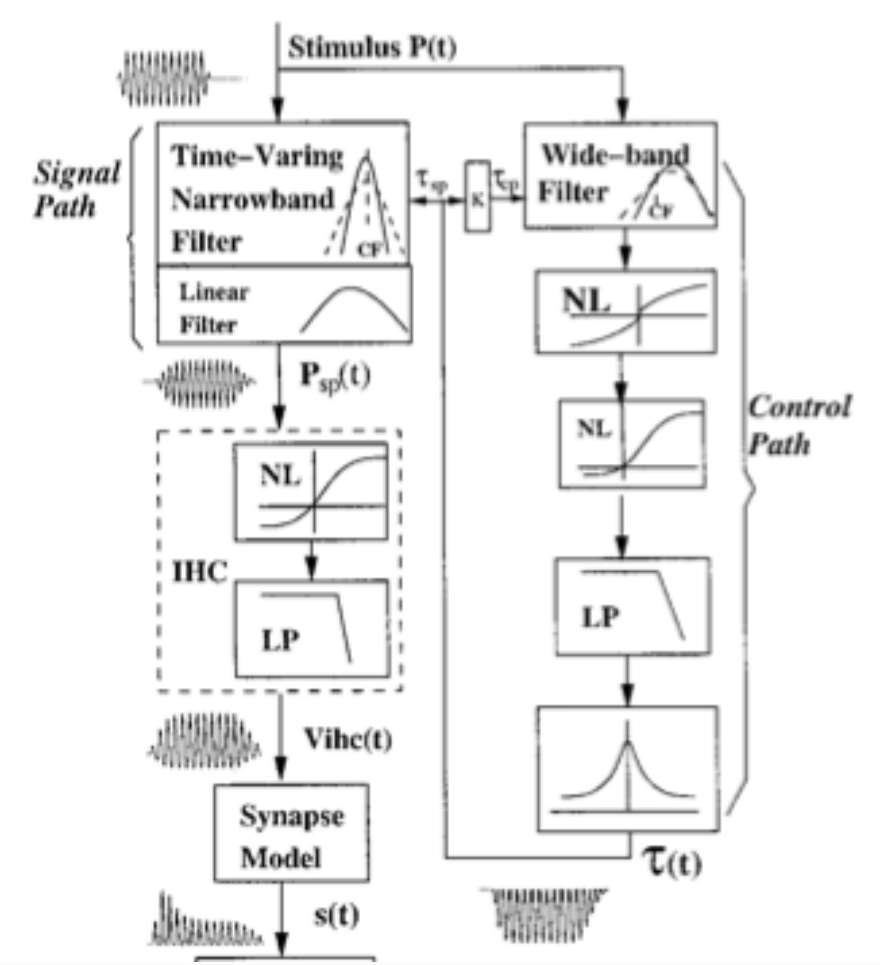
Spectrum "recovered" from MFCC:



Comments on MFCC representation

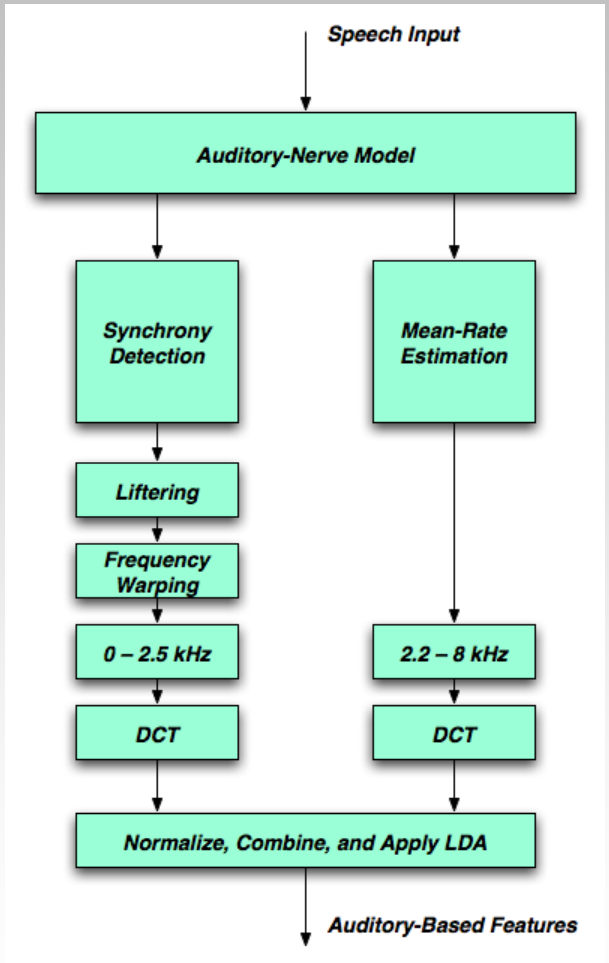
- **It's very "blurry" compared to a wideband spectrogram!**
- **Aspects of auditory processing represented:**
  - Frequency selectivity and spectral bandwidth (but using a constant analysis window duration!)
    - » Wavelet schemes exploit time-frequency resolution better
  - Nonlinear amplitude response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

Physiologically-motivated signal processing: the Zhang-Carney-Zilany model of the periphery



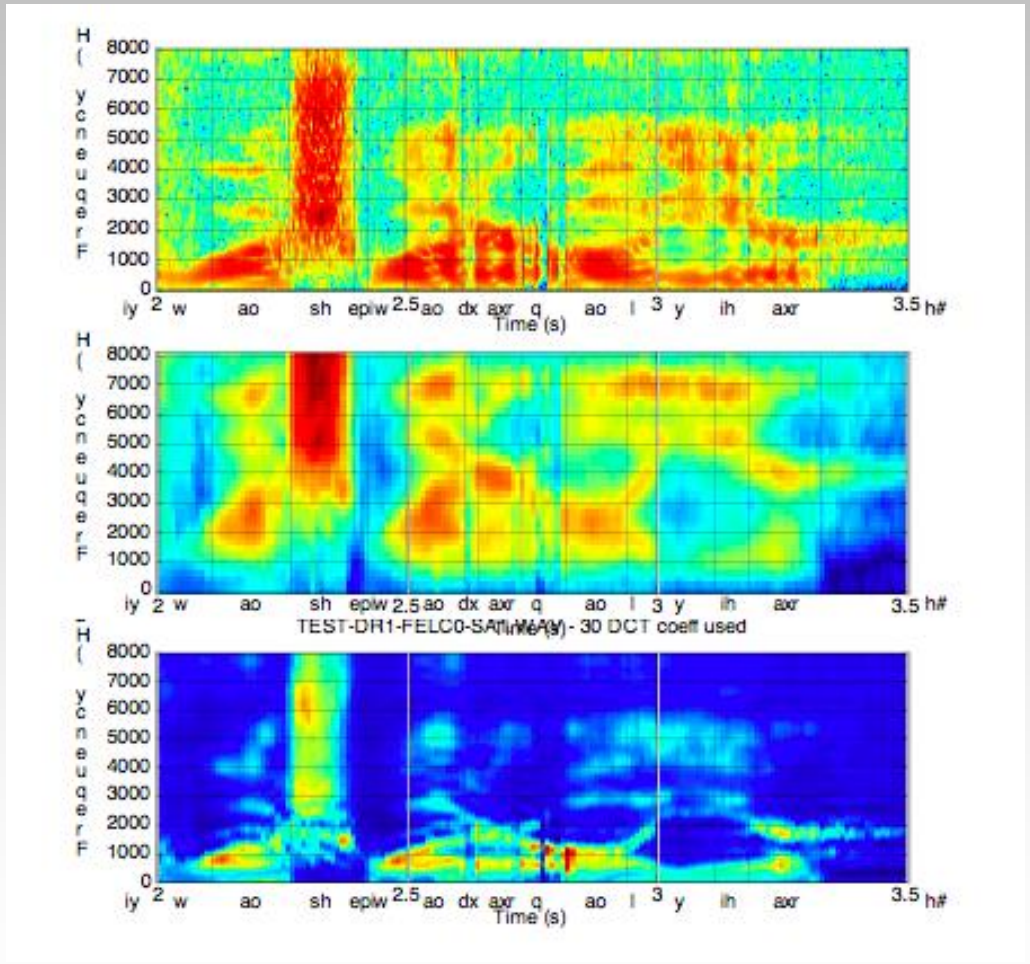
- We used the "synapse output" as the basis for further processing

An early evaluation by Kim et al. (Interspeech 2006)

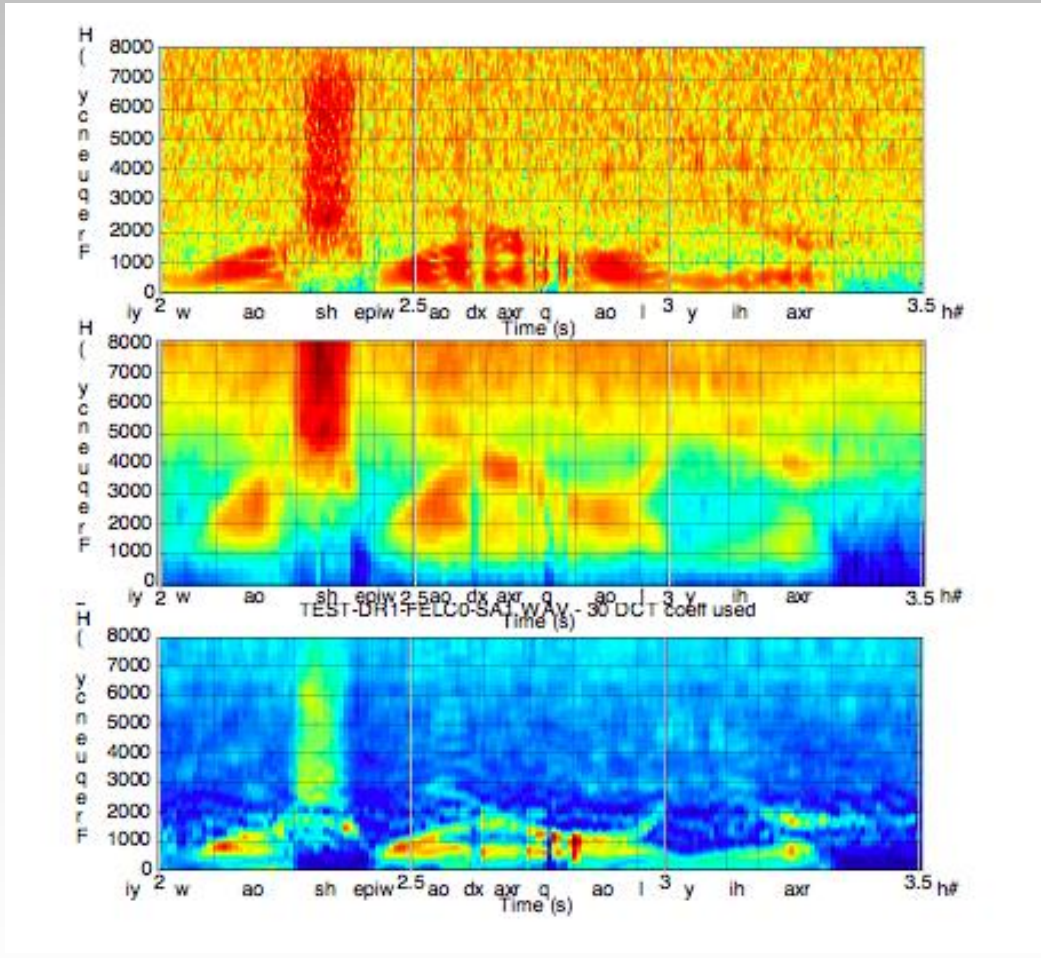


- Synchrony response is smeared across frequency to remove pitch effects
- Higher frequencies represented by mean rate of firing
- Synchrony and mean rate combined additively
- Much more processing than MFCCs, but will simplify if results are useful

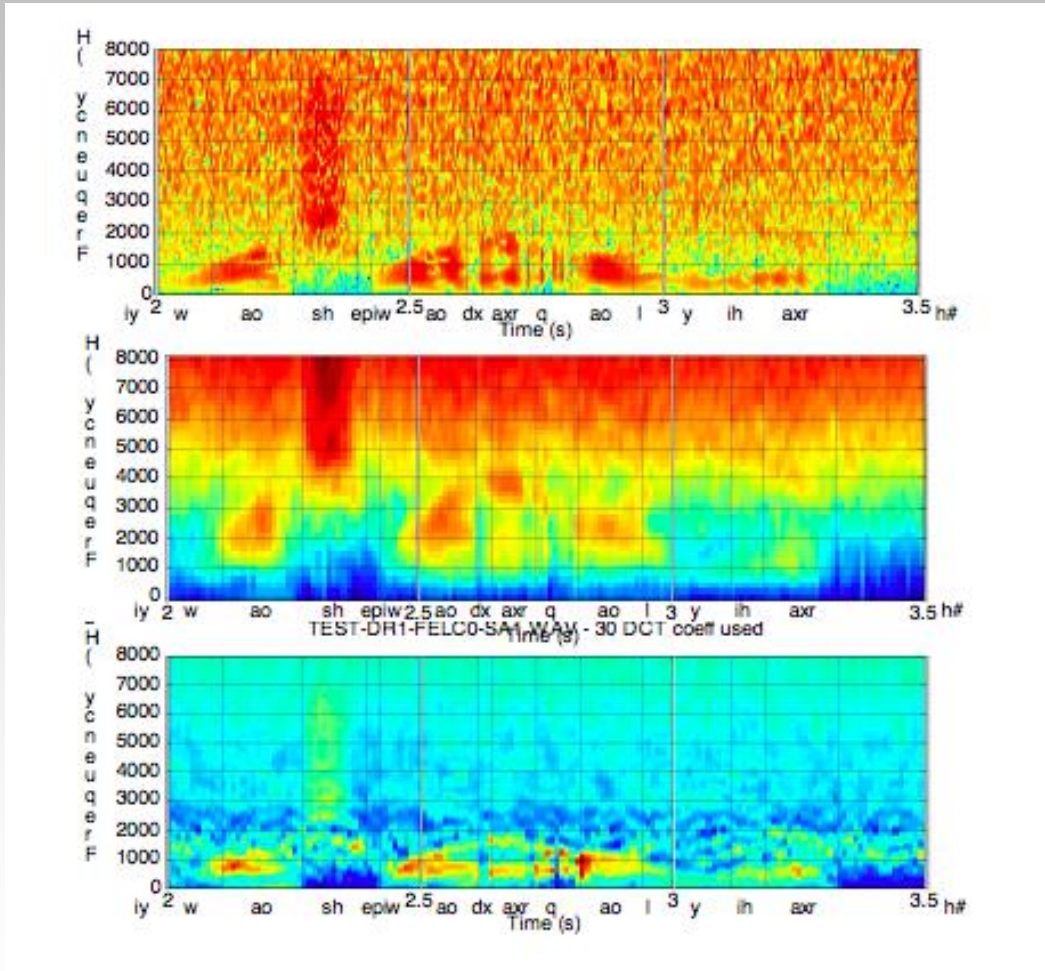
Comparing auditory processing with cepstral analysis: clean speech



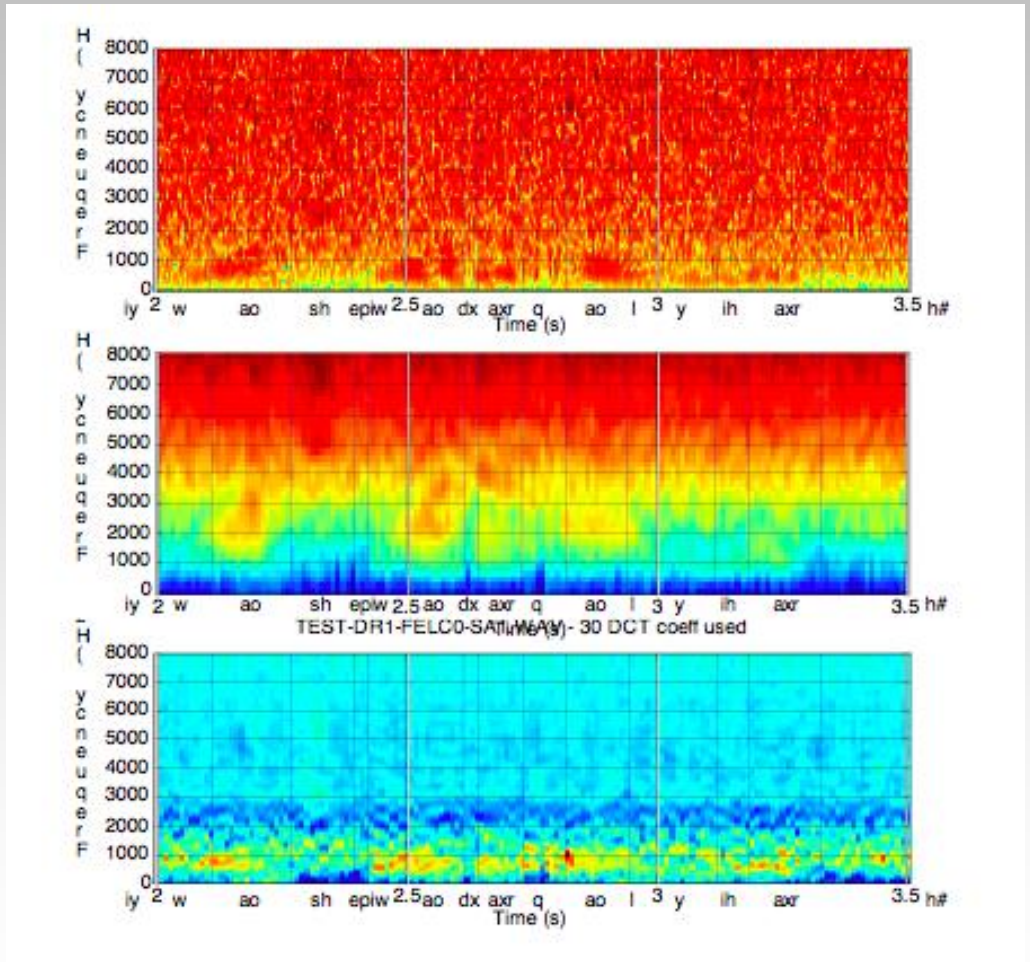
Comparing auditory processing with cepstral analysis: 20-dB SNR



Comparing auditory processing with cepstral analysis: 10-dB SNR

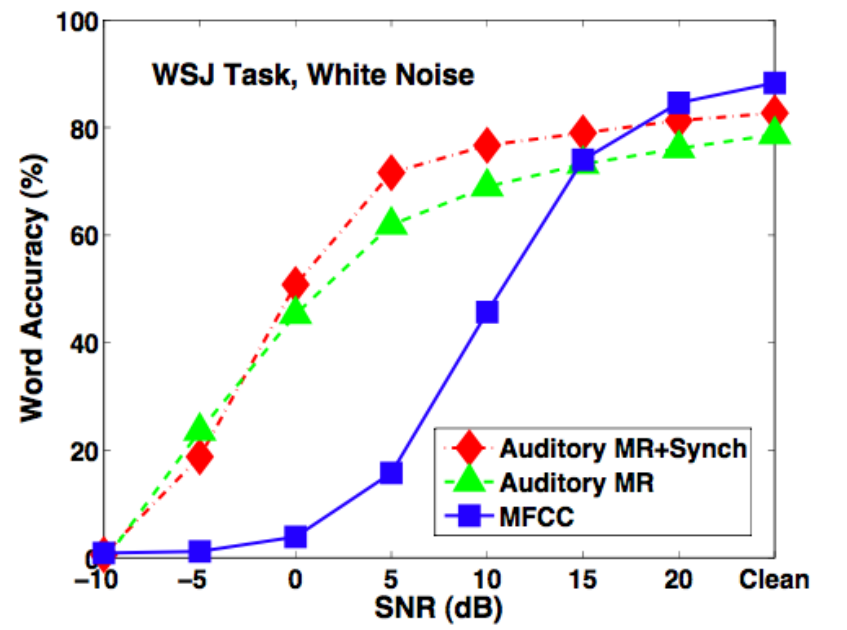


Comparing auditory processing with cepstral analysis: 0-dB SNR

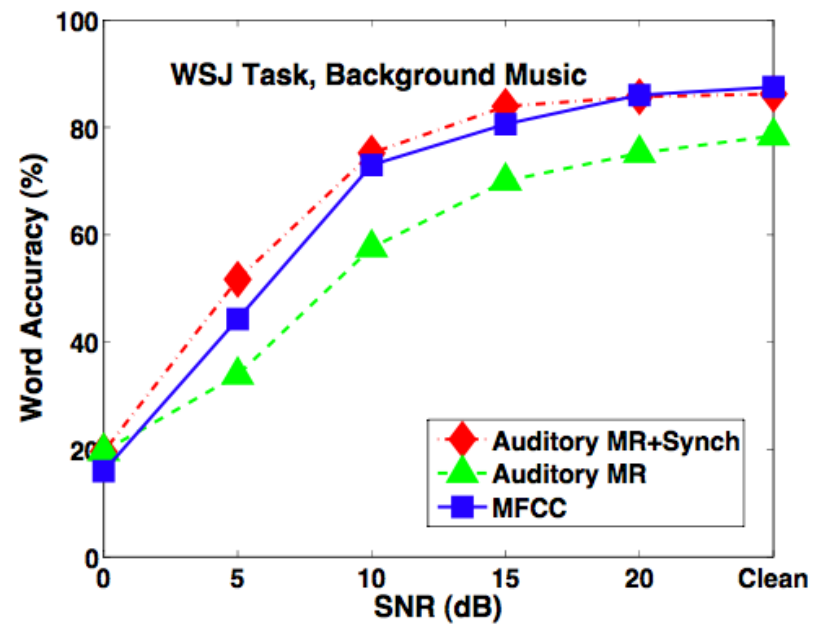


Auditory processing is more effective than MFCCs at low SNRs, especially in white noise

Accuracy in background noise:



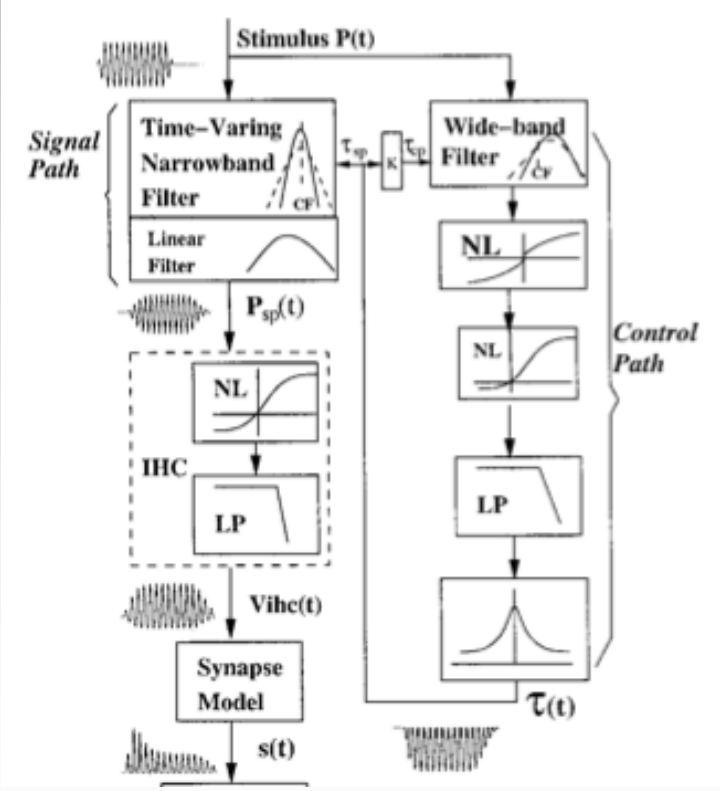
Accuracy in background music:



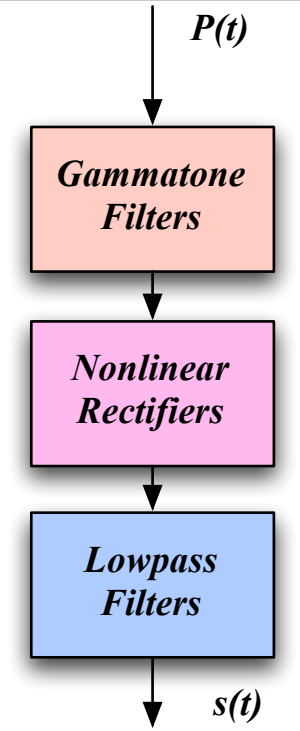
[Results from Kim et al., Interspeech 2006]

But do auditory models really need to be so complex?

- Model of Zhang et al. 2001:

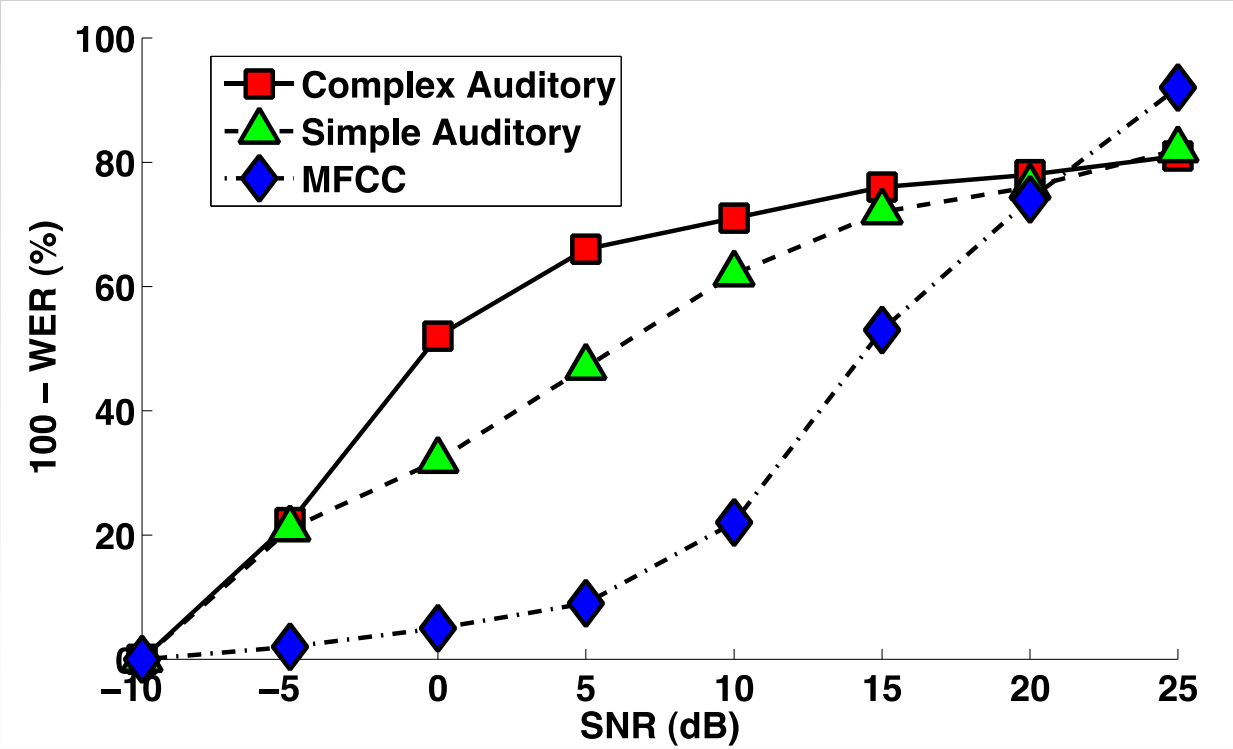


A much simpler model:



Comparing simple and complex auditory models

- Comparing MFCC processing, a trivial (filter–rectify–compress) auditory model, and the full Carney-Zhang model:



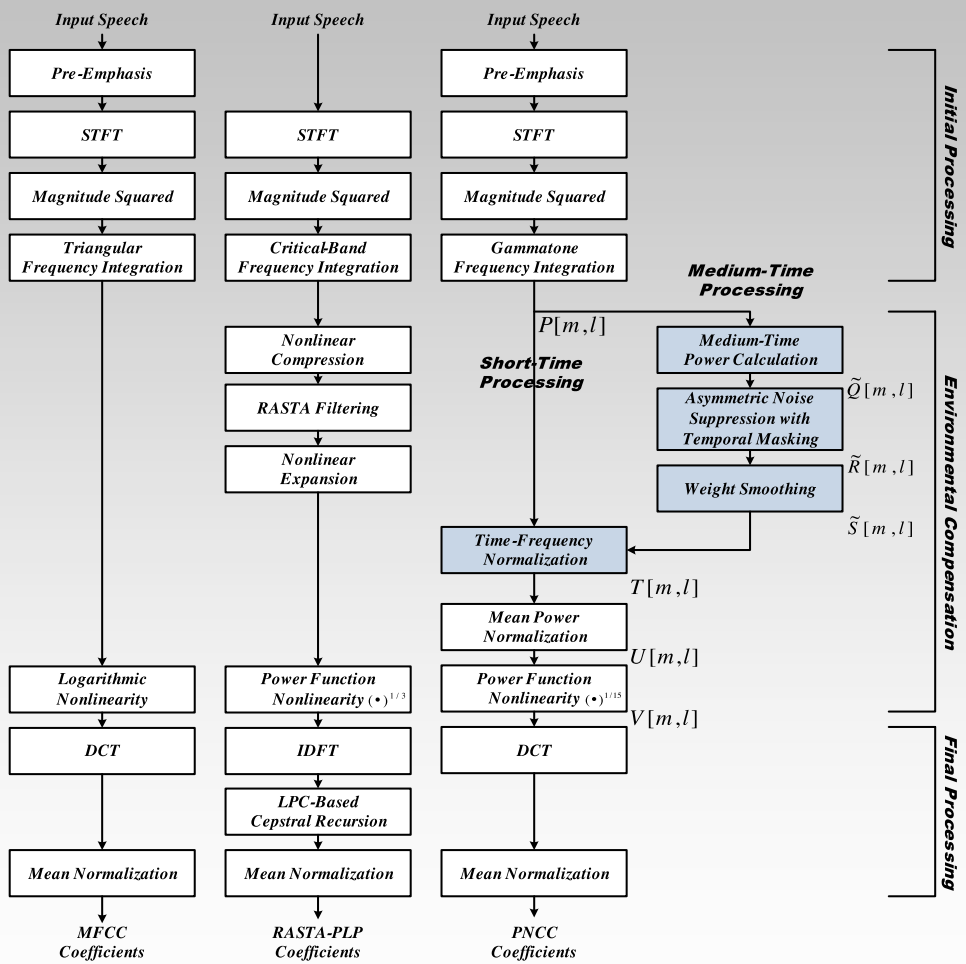
Aspects of auditory processing we have found to be important in improving WER in noise

- The shape of the peripheral filters
- The shape of the auditory nonlinearity
- The use of “medium-time” analysis for noise and reverberation compensation
- The use of nonlinear filtering to obtain noise suppression and general separation of speechlike from non-speechlike signals (a form of modulation filtering)
- The use of nonlinear approaches to effect onset enhancement
- Binaural processing for further enhancement of target signals

PNCC processing (Kim and Stern, 2010,2014)

- **A pragmatic implementation of a number of the principles described:**
  - Gammatone filterbanks
  - Nonlinearity shaped to follow auditory processing
  - “Medium-time” environmental compensation using nonlinearity cepstral highpass filtering in each channel
  - Enhancement of envelope onsets
  - Computationally efficient implementation

PNCC: an integrated front end based on auditory processing



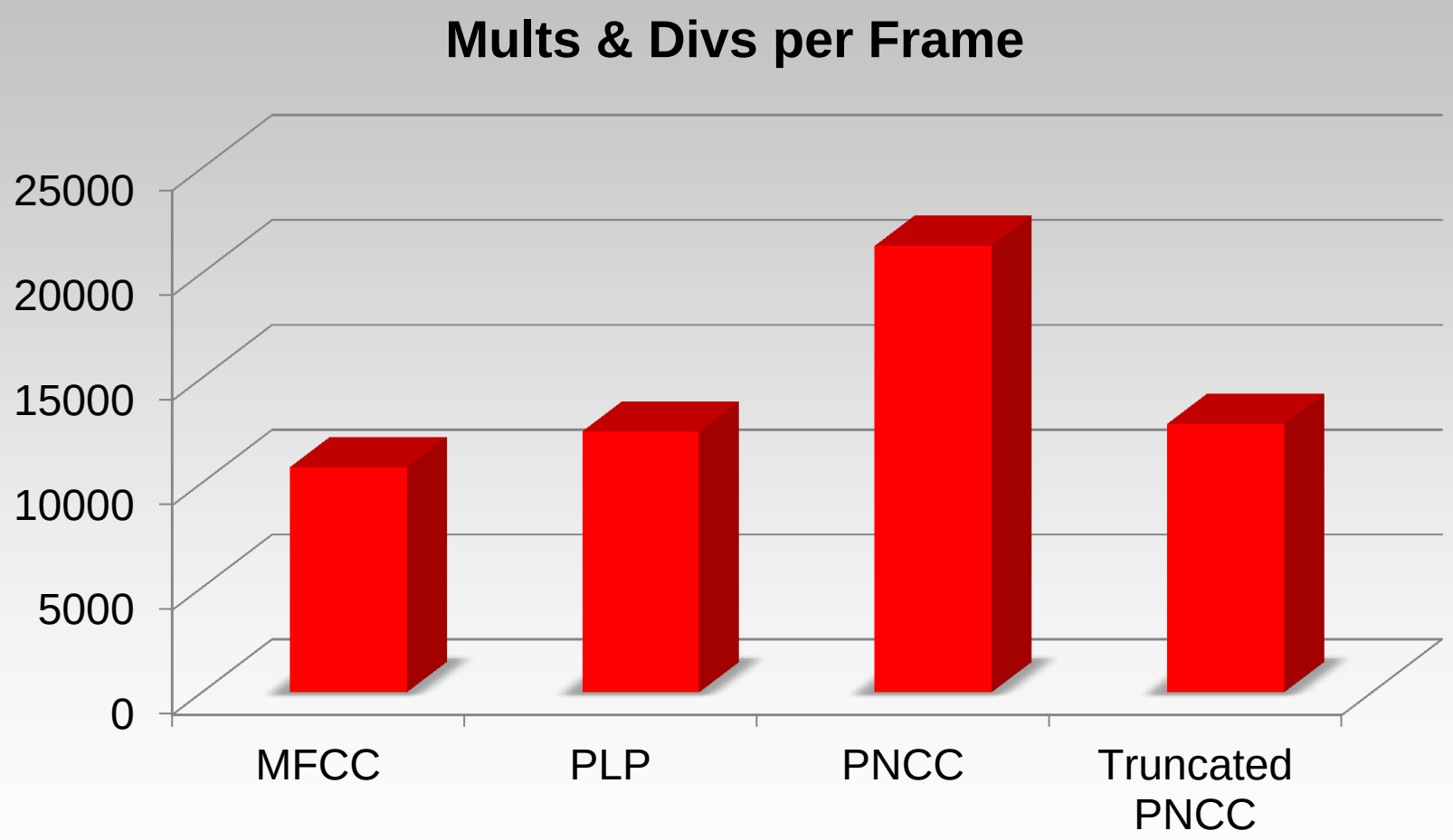
Initial processing

Environmental compensation

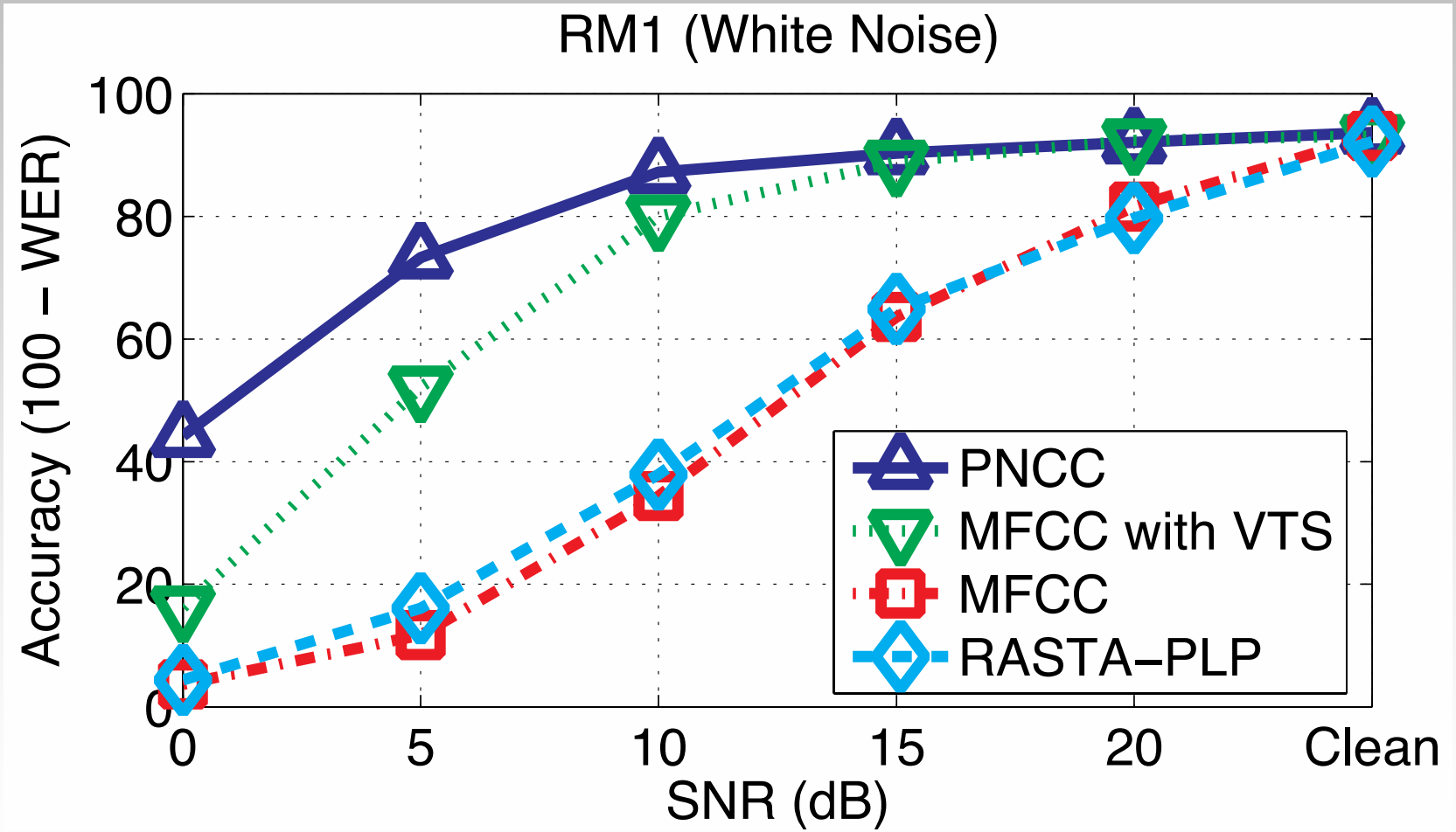
Final processing

MFCC RASTA-PLP PNCC

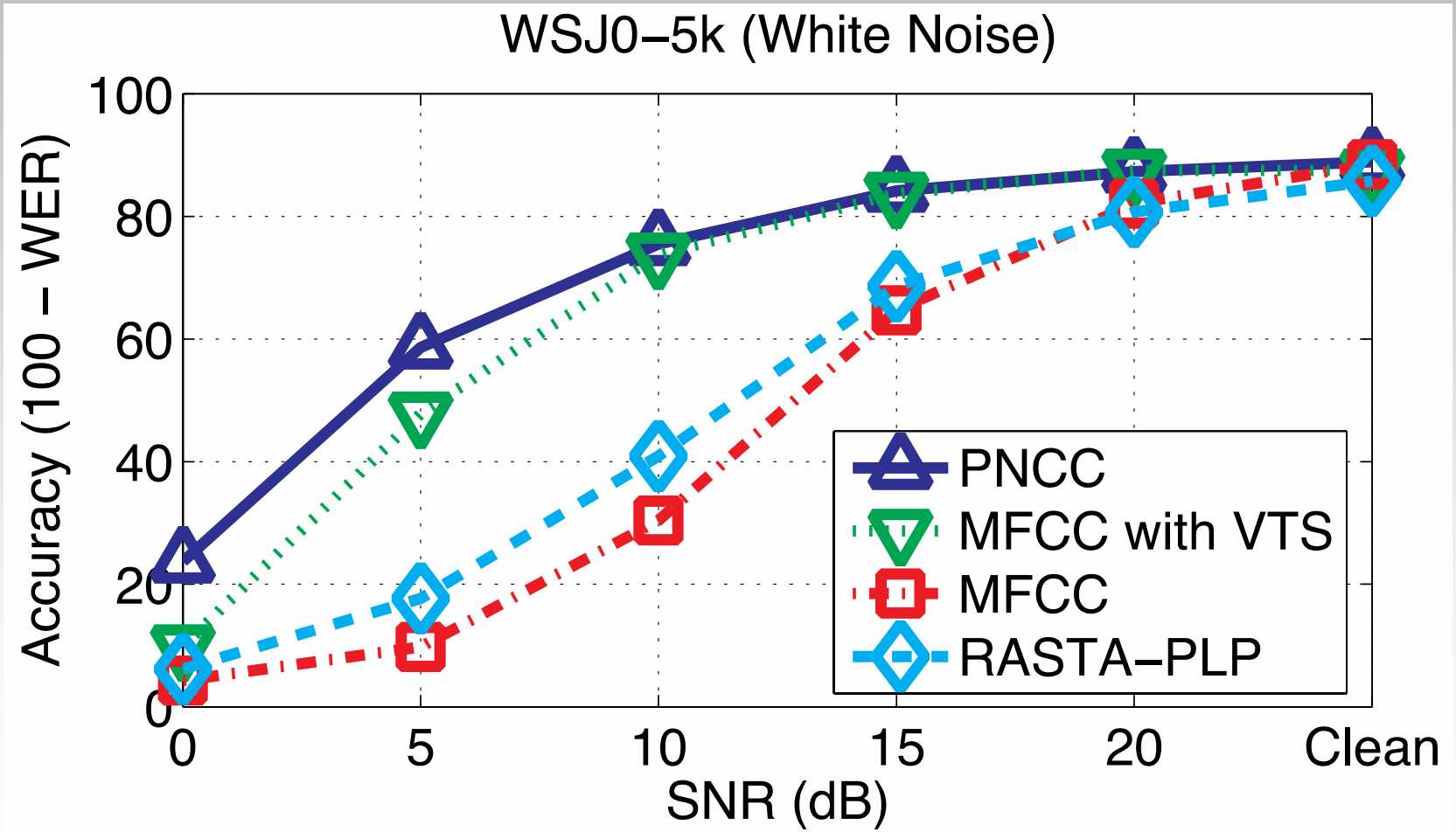
Computational complexity of front ends



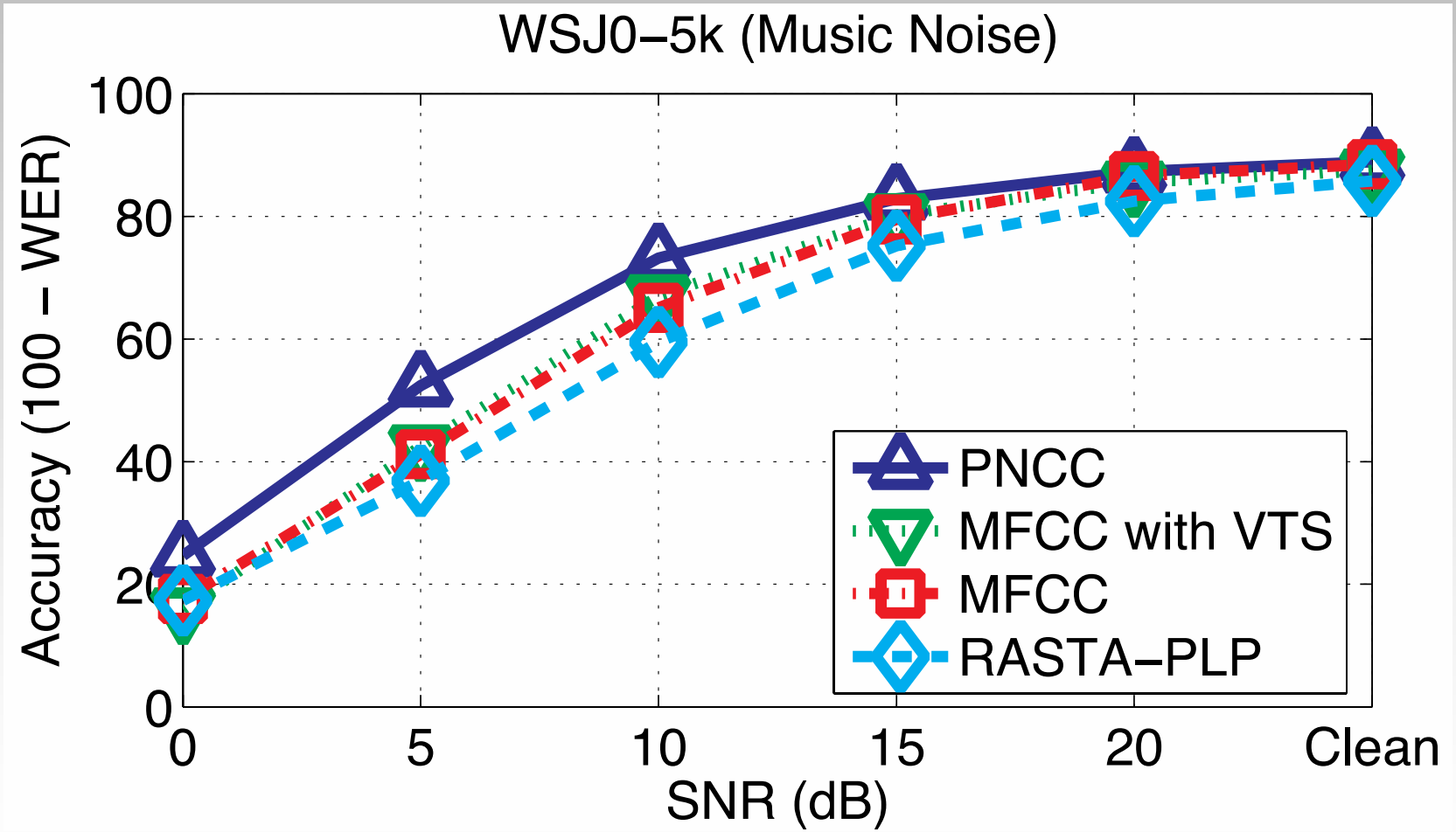
Performance of PNCC in white noise (RM)



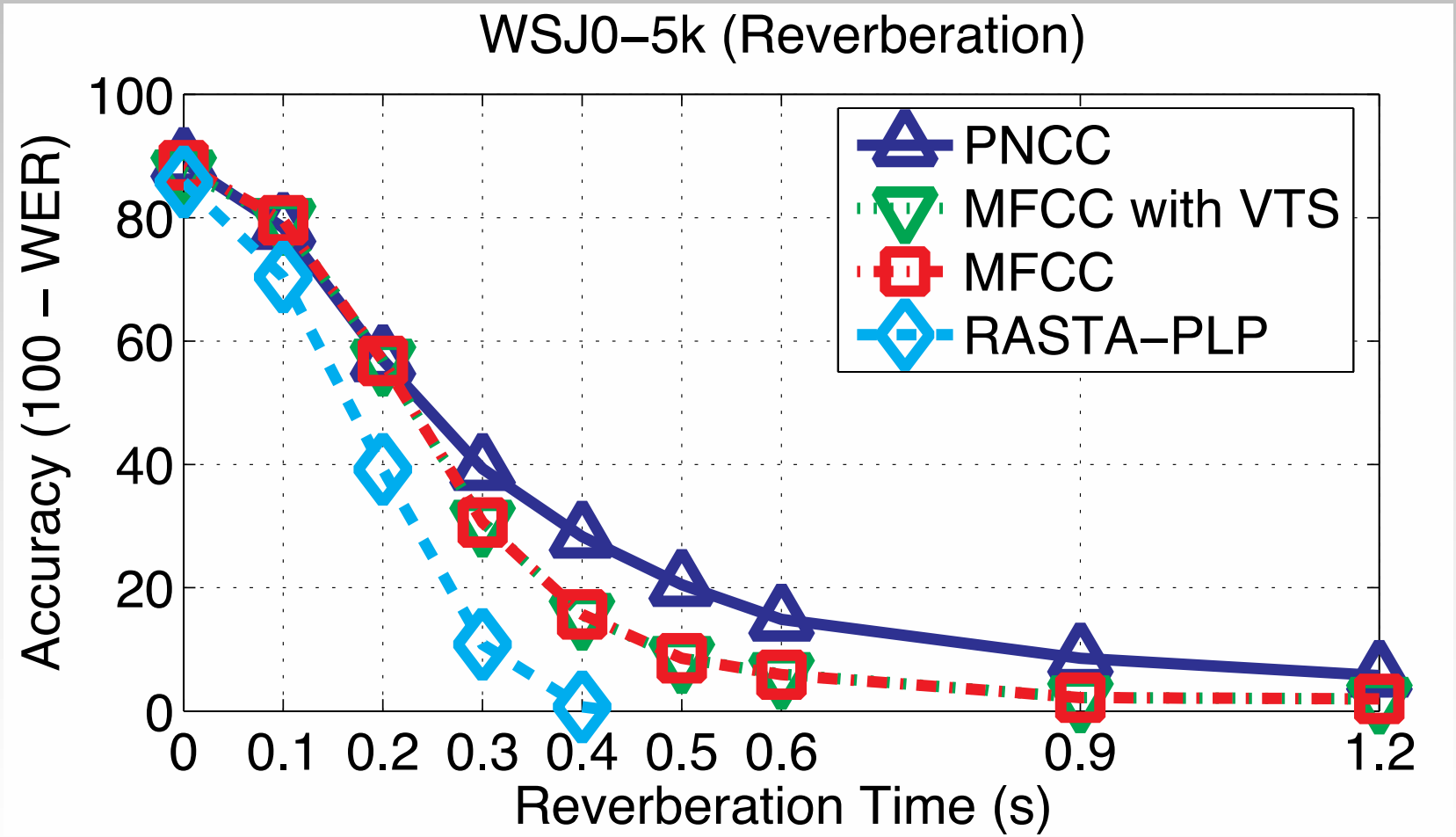
Performance of PNCC in white noise (WSJ)



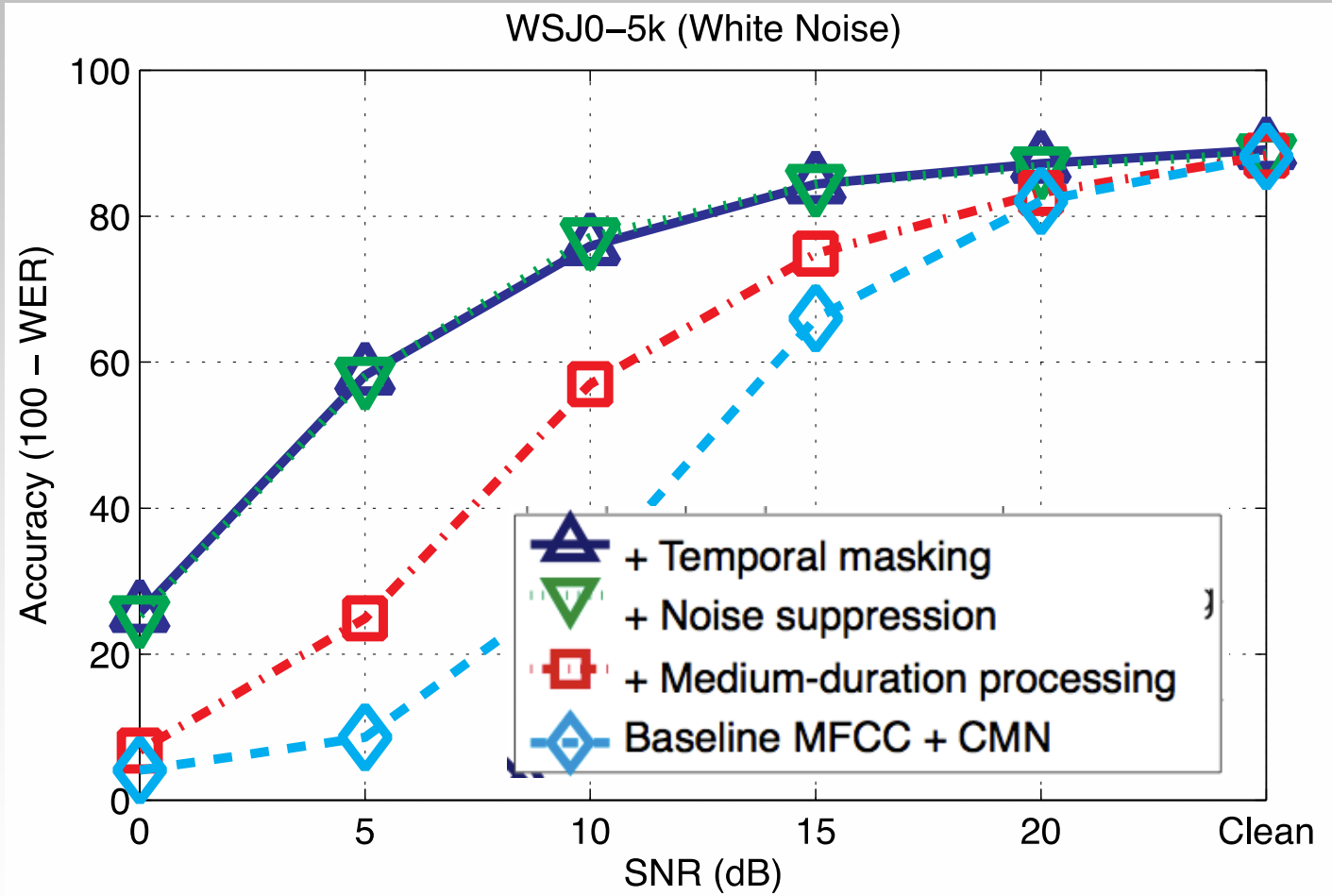
Performance of PNCC in background music



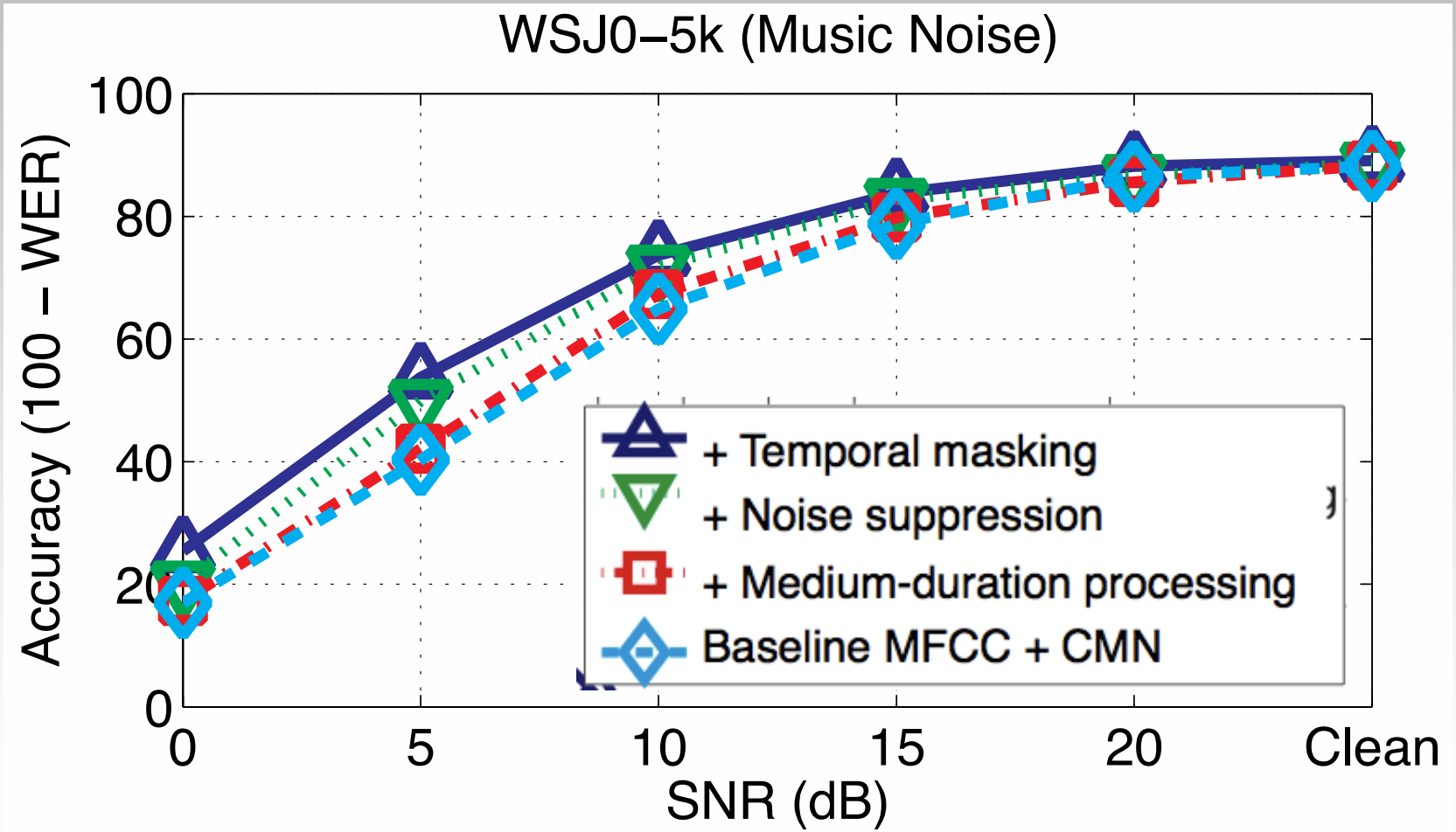
Performance of PNCC in reverberation



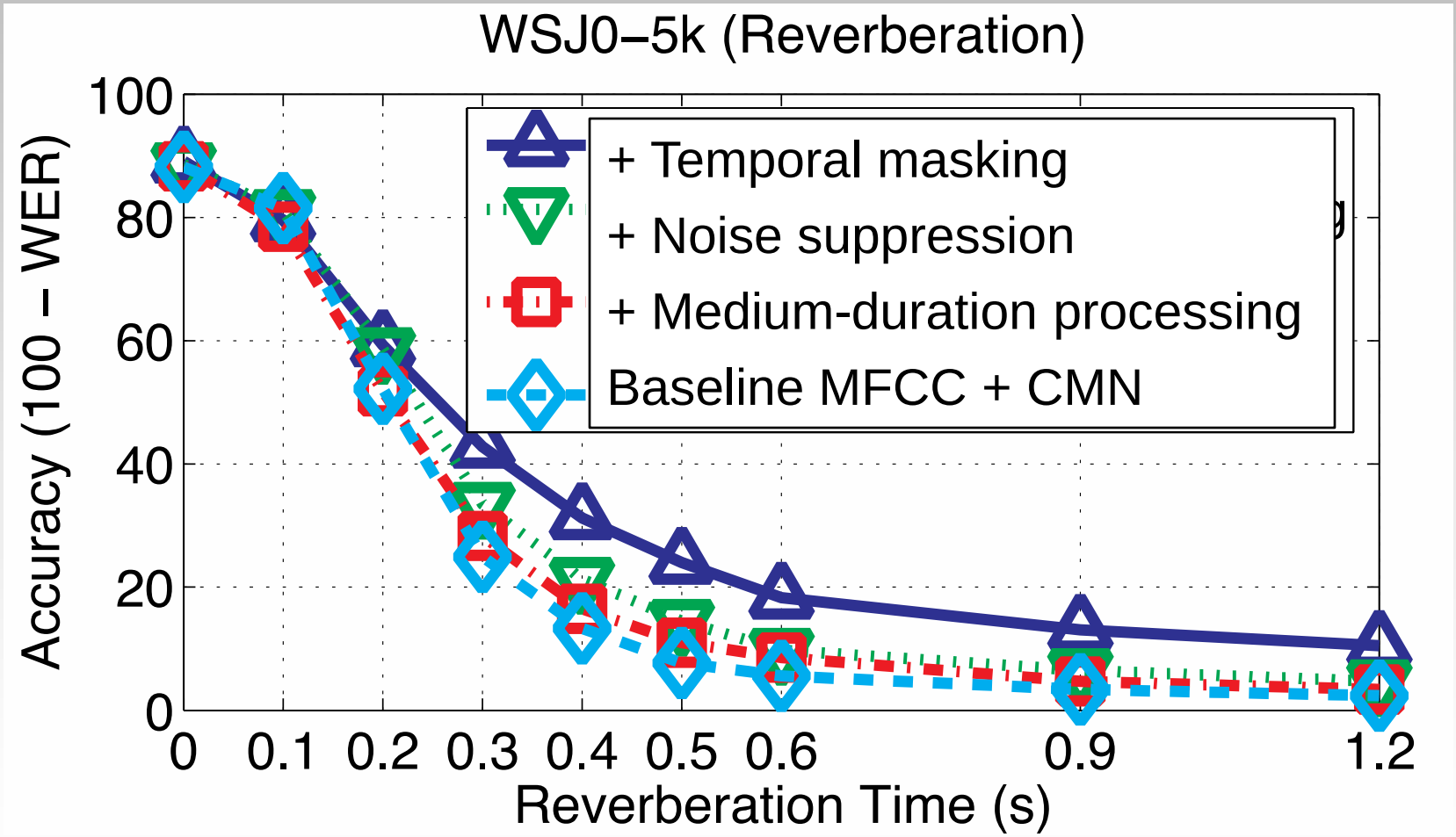
Contributions of PNCC components: white noise (WSJ)



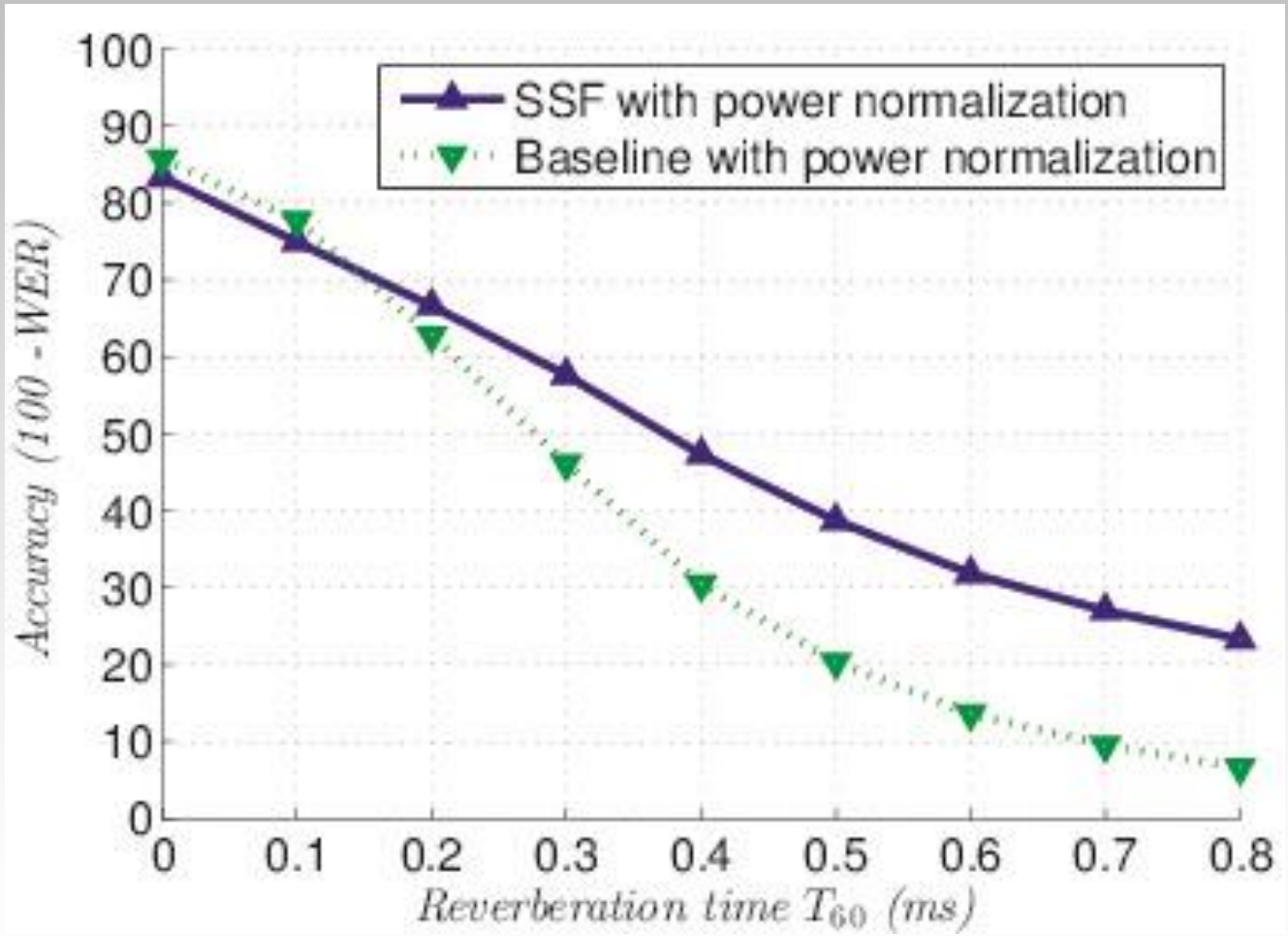
Contributions of PNCC components: background music (WSJ)



Contributions of PNCC components: reverberation (WSJ)



PNCC and SSF @Google



Summary: auditory processing

- **Knowledge of the auditory system will improve ASR accuracy. Important aspects include:**
  - Consideration of filter shapes
  - Consideration of rate-intensity function
  - Onset enhancement
  - Nonlinear modulation filtering
  - Temporal suppression

General summary: Robust recognition in the 21st century

- **Low SNRs, reverberated speech, speech maskers, and music maskers are difficult challenges for robust speech recognition**
- **Robustness algorithms based on classical statistical estimation fail in the presence of transient degradation**
- **Some more recent techniques that can be effective:**
  - Missing-feature reconstruction
  - Microphone array processing in several forms
  - Processing motivated by monaural and binaural physiology and perception
- **More information about this work may be found at**

http://www.cs.cmu.edu/~robust