Applying Models of Auditory Processing to Automatic Speech Recognition: Promise and Progress

Richard Stern
(with Chanwoo Kim, Yu-Hsiang Chiu, and others)

Department of Electrical and Computer Engineering
and Language Technologies Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

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

2014 Frederick Jelinek Memorial Workshop on
Meaning Representations in Language and Speech Processing
Prague, Czech Republic
July 16, 2014

Introduction – auditory processing and automatic speech recognition

- **I was originally trained in auditory perception, and my original work was in binaural hearing**
- **Over the past 20-25 years, I have been spending the bulk of my time trying to improve the accuracy of automatic speech recognition systems in difficult acoustical environments**
- **In this talk I would like to discuss some of the ways in my group (and many others) have been attempting to apply knowledge of auditory perception to improve ASR accuracy**
  - Comment: approaches can be more or less faithful to physiology and psychophysics

The big questions ….

- How can knowledge of auditory physiology and perception improve speech recognition accuracy?
- Can speech recognition results tell us anything we don't already know about auditory processing?
- What aspects of the processing are most valuable for robust feature extraction?

Two historical notes

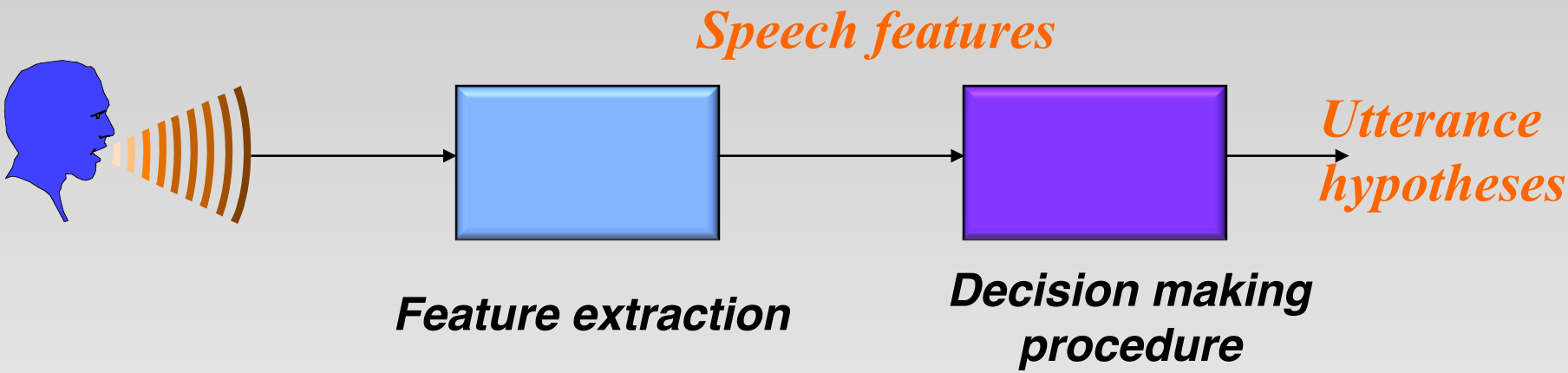
- **Everything is changing with deep learning**
  - Is there a role for “traditional” robust speech technologies?

- **Knowledge-based versus statistically-based processing**

So what I will do is ….

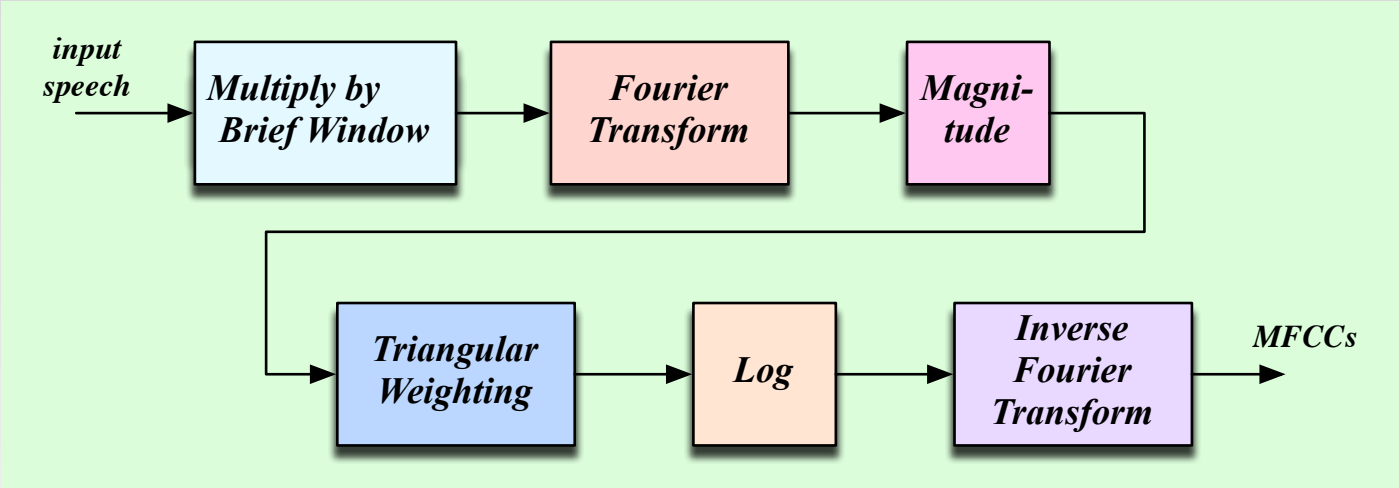
- **Briefly review some of the major physiological and psychophysical results that motivate the models**
- **Briefly review and discuss the major "classical" auditory models of the 1980s**
  - Seneff, Lyon, and Ghitza
- **Review some of the major new trends in today's models**
- **Talk about some representative issues that have driven work as of late at CMU and what we have learned from them**

Speech recognition as pattern classification



- **Major functional components:**
  - Signal processing to extract features from speech waveforms
  - Comparison of features to pre-stored representations
- **Important design choices:**
  - Choice of features
  - Specific method of comparing features to stored representations

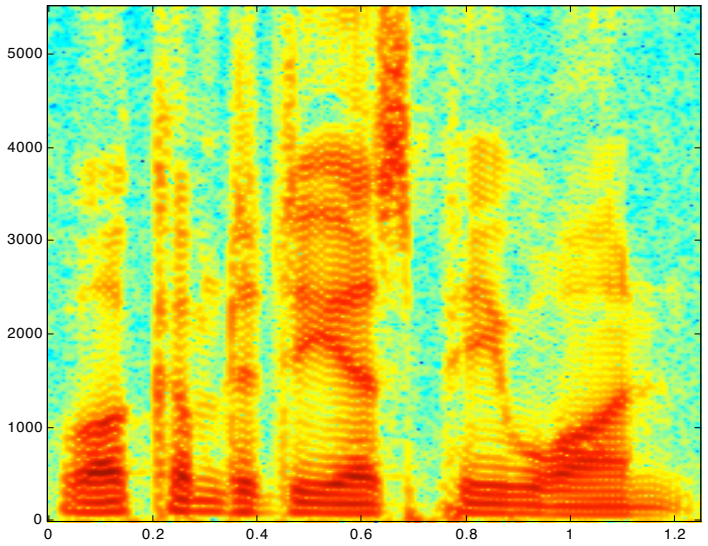
Default signal processing: Mel frequency cepstral coefficients (MFCCs)



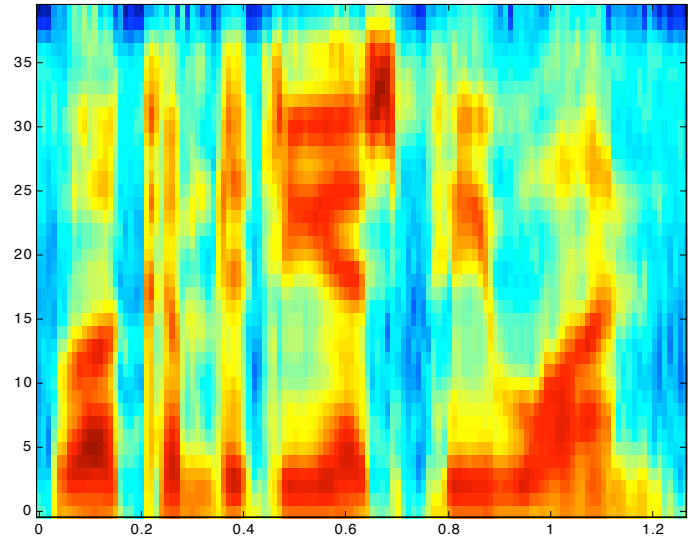
Comment: 20-ms time slices are modeled by smoothed spectra, with attention paid to auditory frequency selectivity

What the speech recognizer sees ….

An original spectrogram:



Spectrum “recovered” from MFCC:

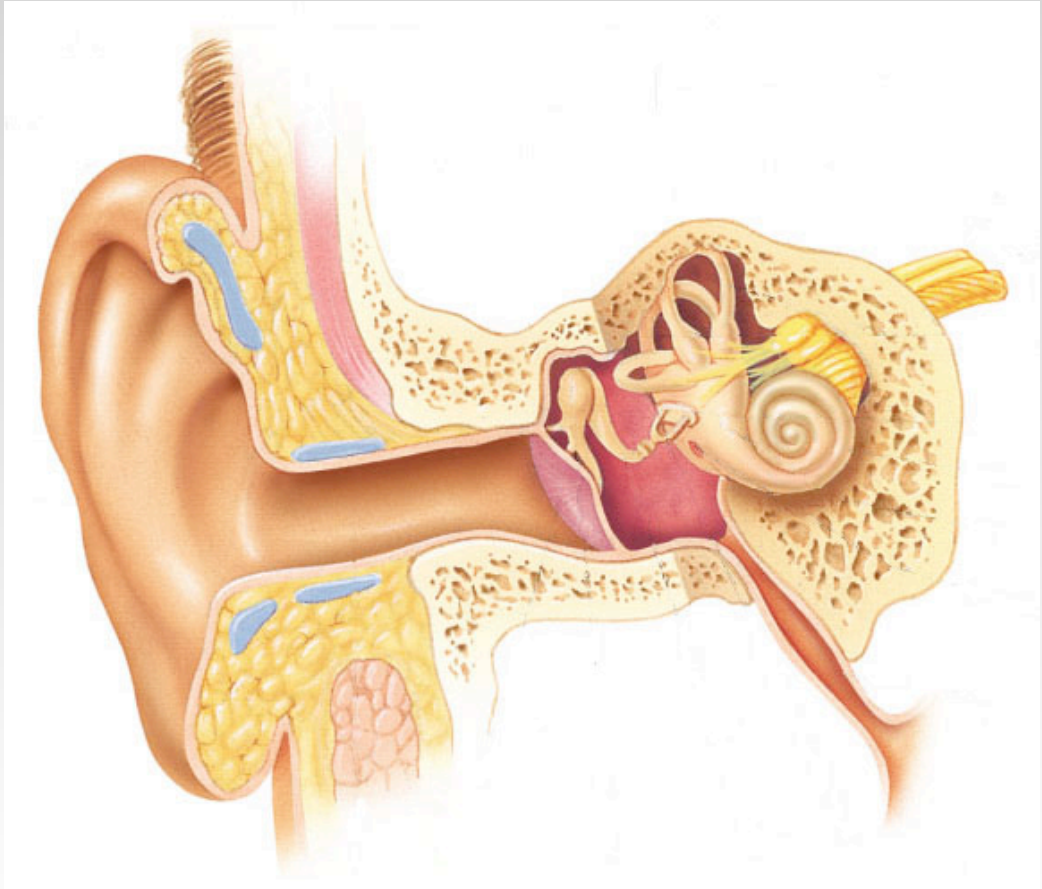


Comments on the MFCC representation

- **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
- **Aspects of auditory processing NOT represented:**
  - Detailed timing structure
  - Lateral suppression
  - Enhancement of temporal contrast
  - Other auditory nonlinearities

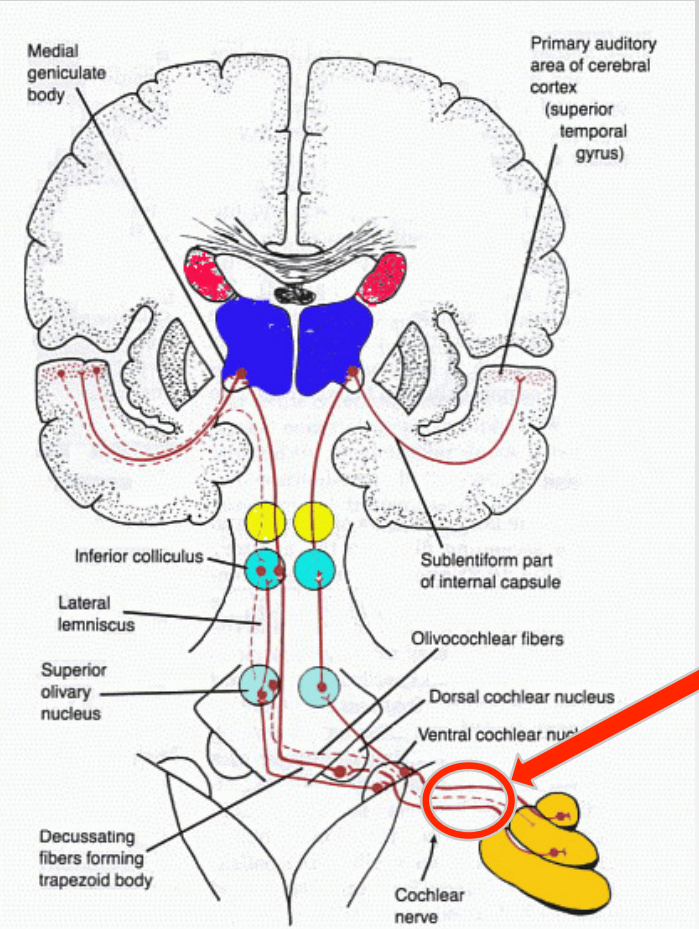
Basic auditory anatomy

- **Structures involved in auditory processing:**



Excitation along the basilar membrane
(courtesy James Hudspeth, HHMI)

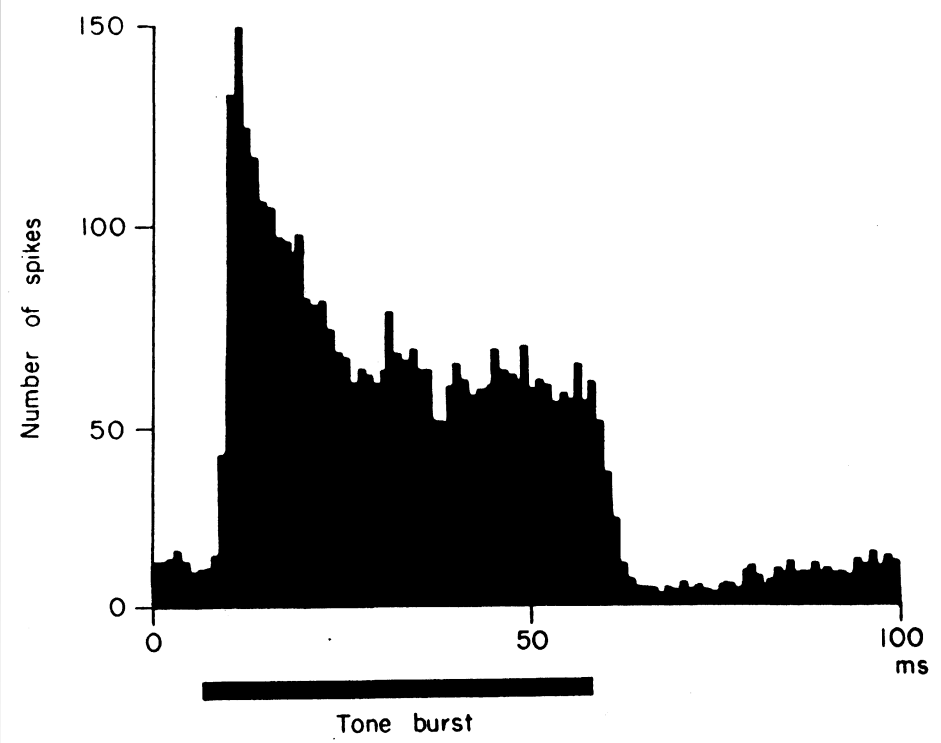
Central auditory pathways



- **There is a lot going on!**
- **For the most part, we only consider the response of the auditory nerve**
  - It is in series with everything else

Transient response of auditory-nerve fibers

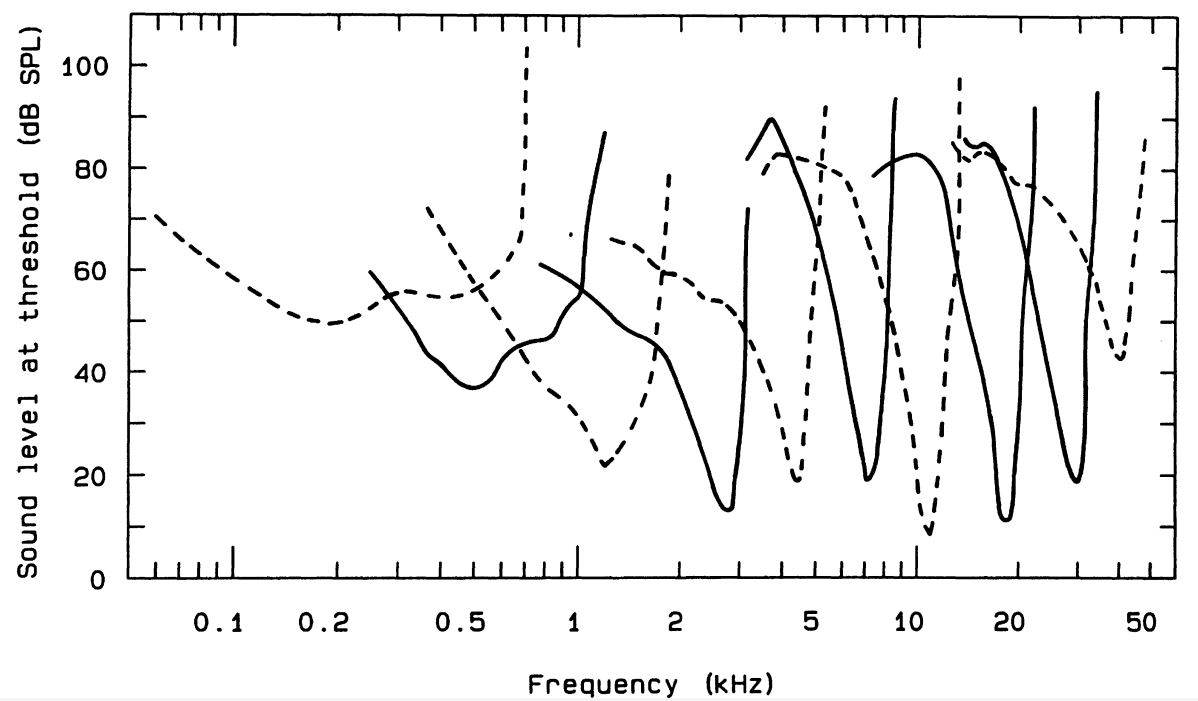
- Histograms of response to tone bursts (Kiang *et al.*, 1965):



Comment: Onsets and offsets produce overshoot

Frequency response of auditory-nerve fibers: tuning curves

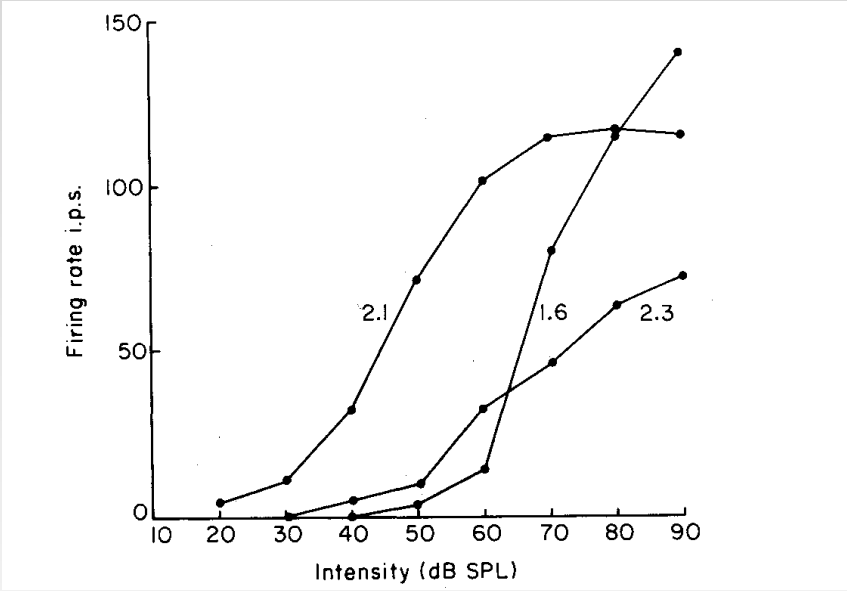
- Threshold level for auditory-nerve response to tones:



- Note dependence of bandwidth on center frequency and asymmetry of response

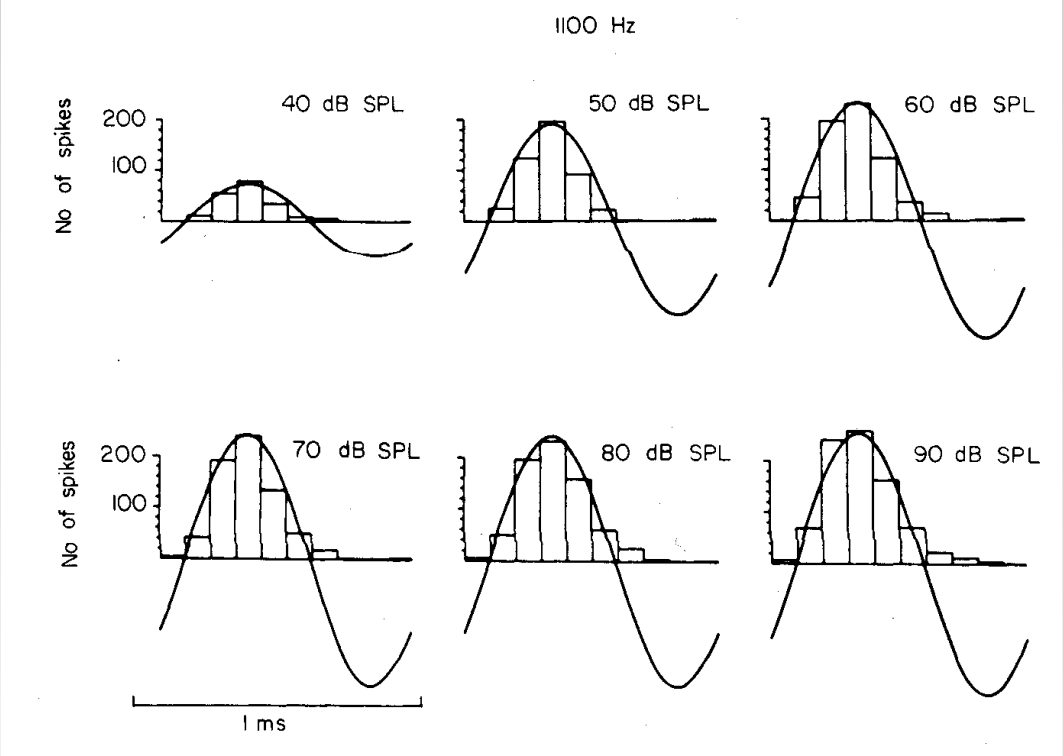
Typical response of auditory-nerve fibers as a function of stimulus level

- **Typical response of auditory-nerve fibers to tones as a function of intensity:**



- **Comment:**
  - Saturation and limited dynamic range

Synchronized auditory-nerve response to low-frequency tones



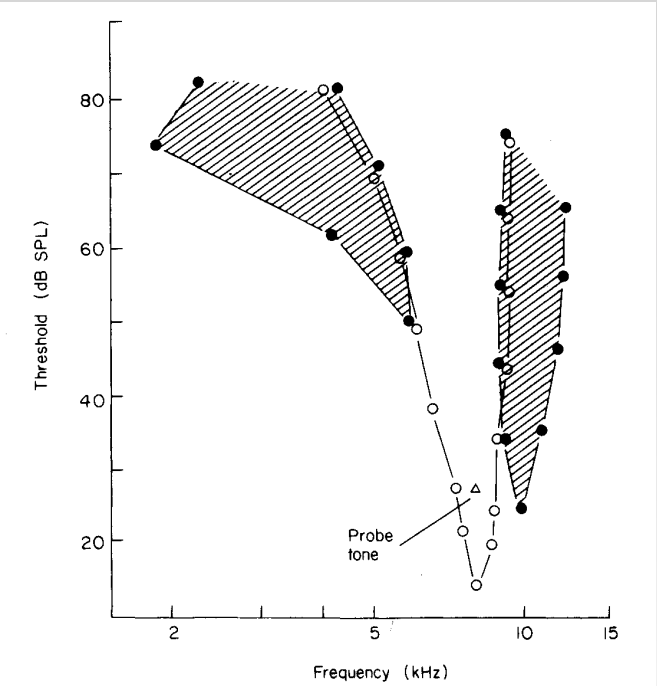
- **Comment: response remains synchronized over a wide range of intensities**

Comments on synchronized auditory response

- Nerve fibers synchronize to fine structure at “low” frequencies, signal envelopes at high frequencies
- Synchrony clearly important for auditory localization
- Synchrony could be important for monaural processing of complex signals as well

Lateral suppression in auditory processing

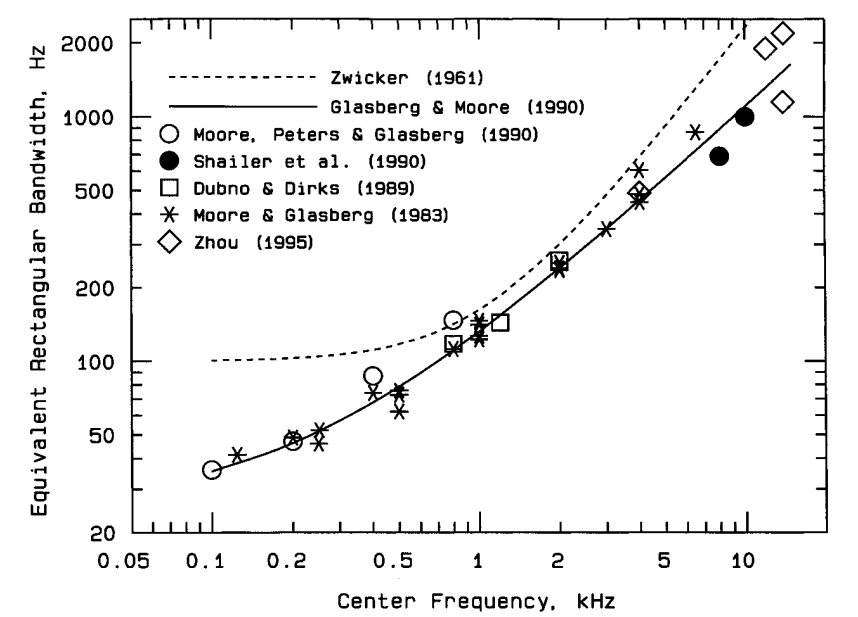
- **Auditory-nerve response to pairs of tones:**



- **Comment: Lateral suppression enhances local contrast in frequency**

Auditory frequency selectivity: critical bands

- **Measurements of psychophysical filter bandwidth by various methods:**



- **Comments:**
  - Bandwidth increases with center frequency
  - Solid curve is “Equivalent Rectangular Bandwidth” (ERB)

Three perceptual auditory frequency scales

Bark scale:
(DE)

$$Bark(f) = \begin{cases} .01f, & 0 \le f < 500 \\ .007f + 1.5, & 500 \le f < 1220 \\ 6\ln(f) - 32.6, & 1220 \le f \end{cases}$$

Mel scale:
(USA)

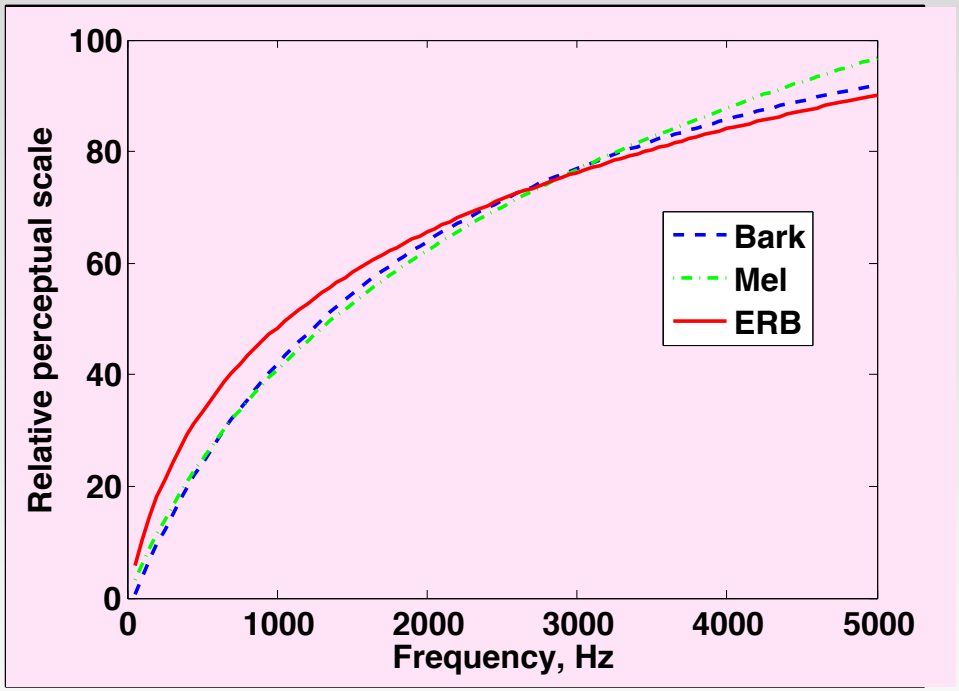
$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

ERB scale:
(UK)

$$ERB(f) = 24.7(4.37f + 1)$$

Comparison of normalized perceptual frequency scales

- Bark scale (in blue), Mel scale (in red), and ERB scale (in green):

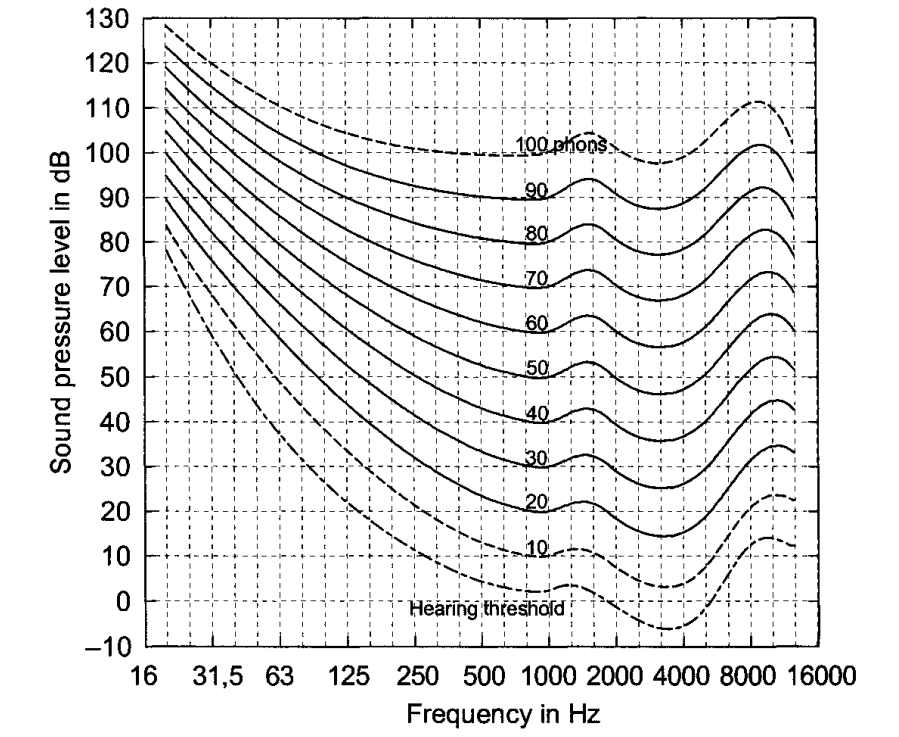


Perceptual masking of adjacent spectro-temporal components

- **Spectral masking:**
  - Intense signals at a given frequency mask adjacent frequencies (asymmetrically)
- **Temporal masking:**
  - Intense signals at a given frequency can mask successive input at that frequency (and to some extent before the masker occurs!)
- **These phenomena are an important part of the auditory models used in perceptual audio coding (such as in creating MP3 files)**

The loudness of sounds

- **Equal loudness contours (Fletcher-Munson curves):**

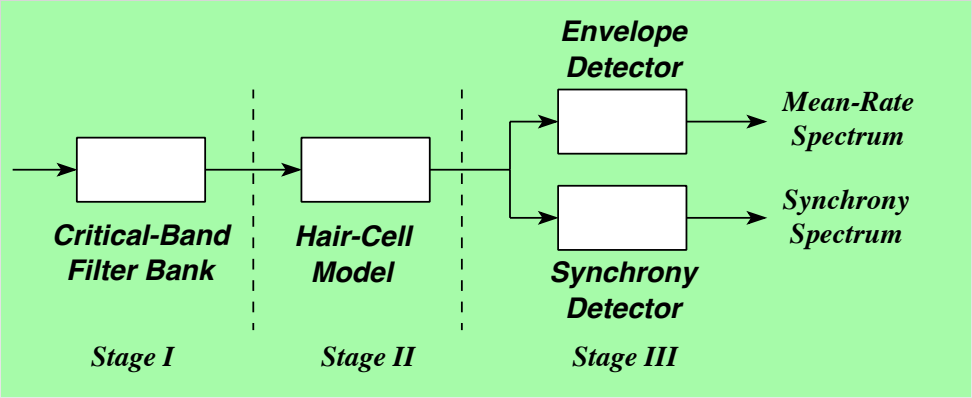


Summary of basic auditory physiology and perception

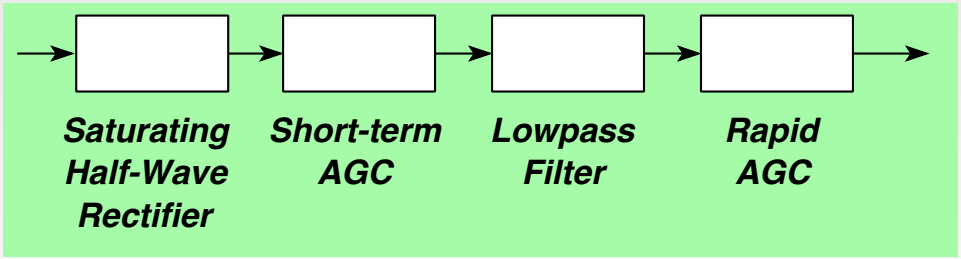
- **Major monaural physiological attributes:**
  - Frequency analysis in parallel channels
  - Preservation of temporal fine structure
  - Limited dynamic range in individual channels
  - Enhancement of temporal contrast (at onsets and offsets)
  - Enhancement of spectral contrast (at adjacent frequencies)
- **Most major physiological attributes have psychophysical correlates**
- **Most physiological and psychophysical effects are not preserved in conventional representations for speech recognition**

Auditory models in the 1980s: the Seneff model

- **Overall model:**



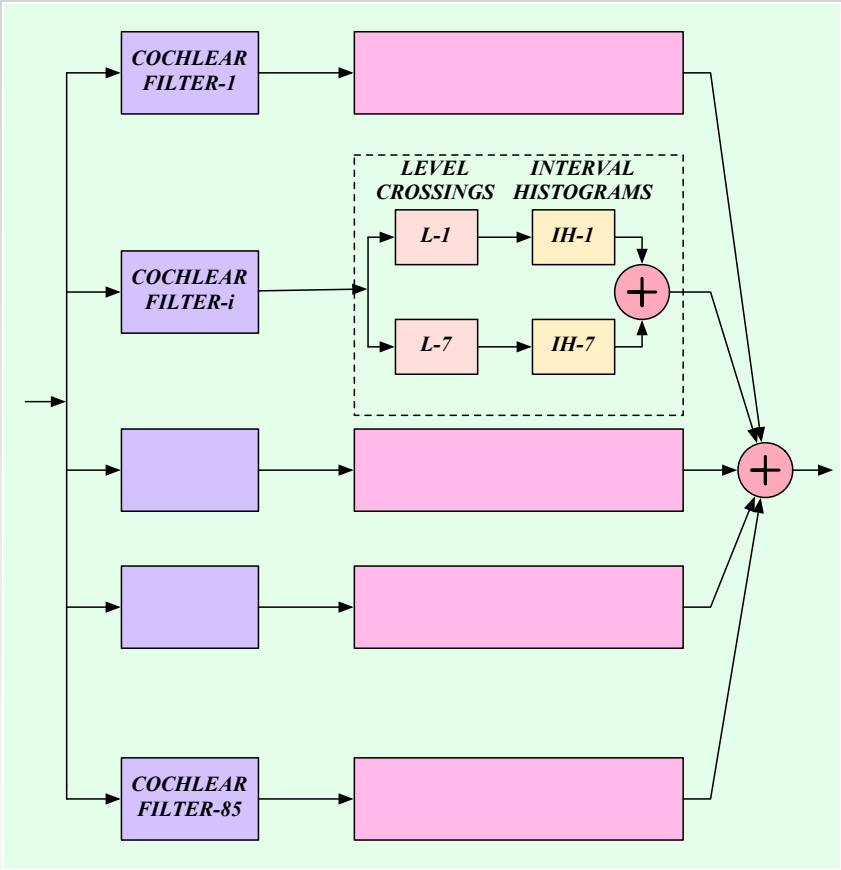
- **Detail of Stage II:**



– An early well-known auditory model

– In addition to mean rate, used "Generalized Synchrony Detector" to extract synchrony

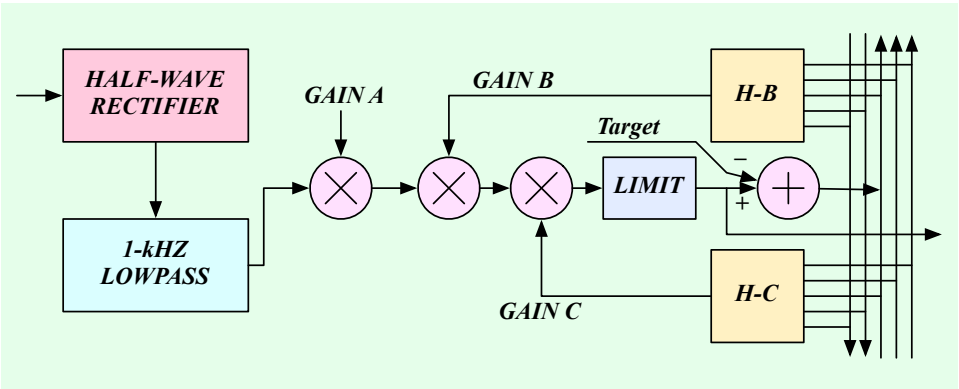
Auditory models in the 1980s: Ghitza's EIH model



– Estimated timing information from ensembles of zero crossings with different thresholds

Auditory models in the 1980s: Lyon's auditory model

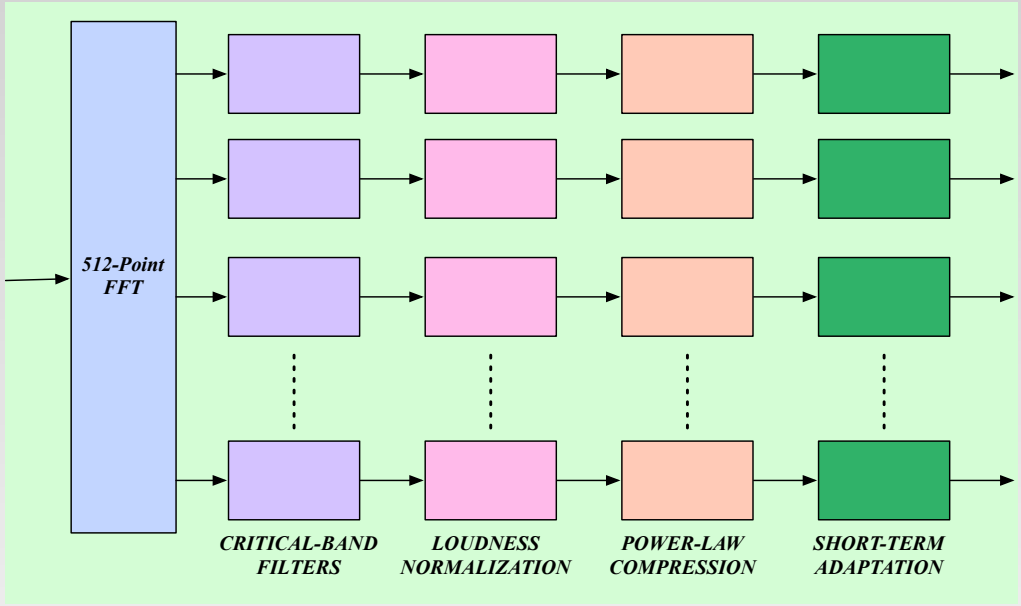
- **Single stage of the Lyon auditory model:**



– Lyon model included nonlinear compression, lateral suppression, temporal effects

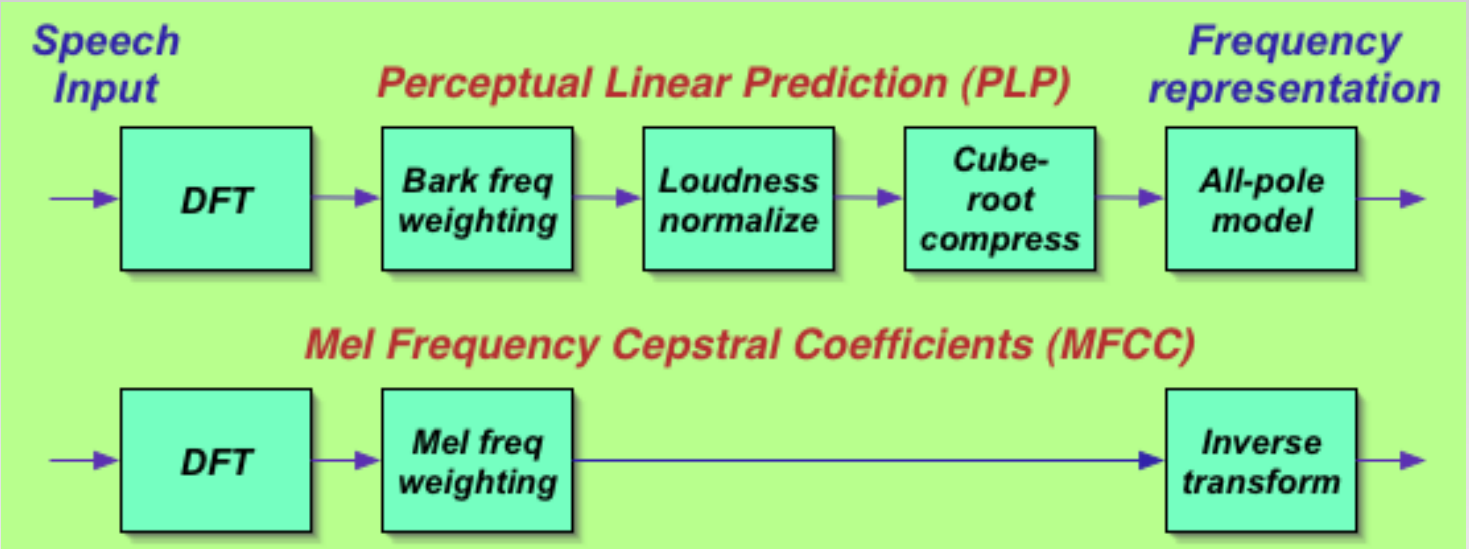
– Also added correlograms (autocorrelation and crosscorrelation of model outputs)

And one more …
Cohen's model (1989)



- Loudness normalization and transient enhancement novel for the time
- Used successfully as part of many IBM systems

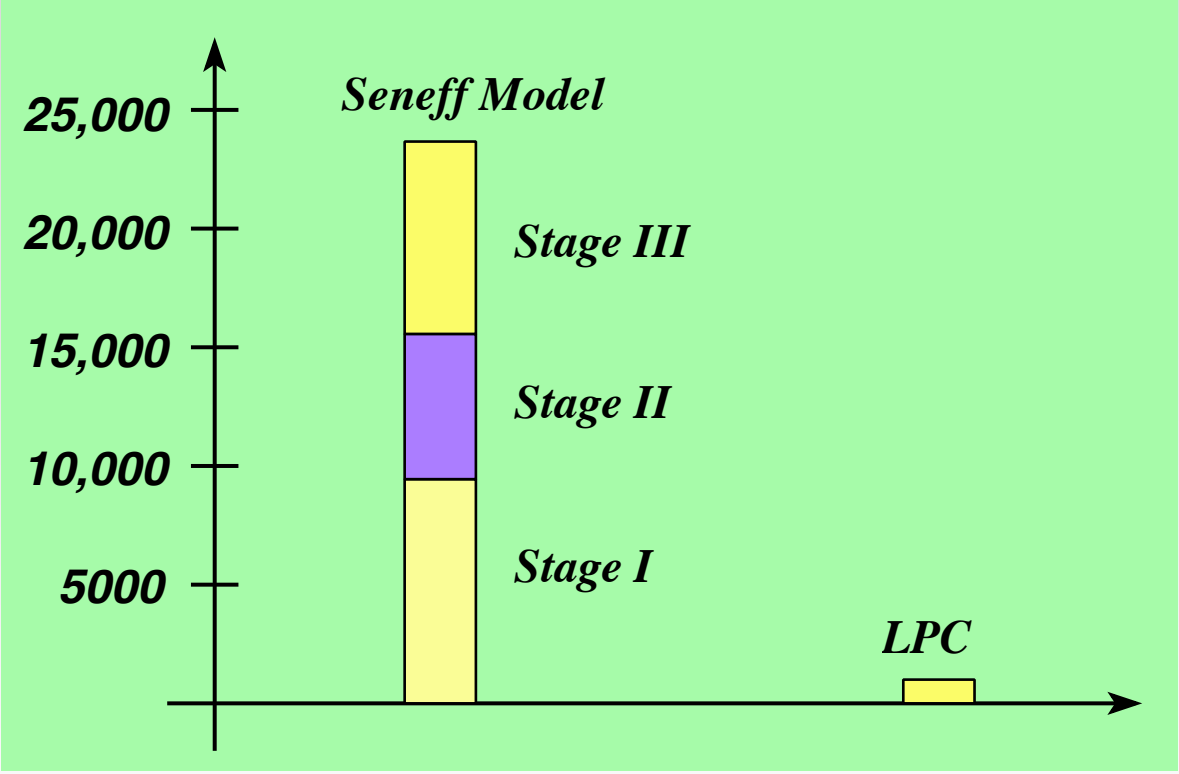
The other standard approach: Perceptual Linear Prediction (PLP, Hermansky '90)



- **Comments**:
  - A pragmatic approach to auditory modeling
  - Pre-emphasis, loudness normalization based on threshold of hearing
  - RASTA enhancement provides cepstral normalization and modulation filtering
  - Widely used with success today

Auditory modeling was expensive: Computational complexity of Seneff model

- **Number of multiplications per ms of speech** (from Ohshima and Stern, 1994):



Summary: early auditory models

- **The models developed in the 1980s included:**
  - "Realistic" auditory filtering
  - "Realistic" auditory nonlnearity
  - Synchrony extraction
  - Lateral suppression
  - Higher order processing through auto-correlation and cross-correlation
- **Every system developer had his or her own idea of what was important**

Evaluation of early auditory models (Ohshima and Stern, 1994)

- **Not much quantitative evaluation actually performed**
- **General trends of results:**
  - Physiological processing did not help much (if at all) for clean speech
  - More substantial improvements observed for degraded input
  - Benefits generally do not exceed what could be achieved with more prosaic approaches (e.g. CDCN/VTS in our case).

Other reasons why work on auditory models subsided in the late 1980s …

- **Failure to obtain a good statistical match between characteristics of features and speech recognition system**
  - Ameliorated by subsequent development of continuous HMMs
- **More pressing need to solve other basic speech recognition problems**

Renaissance in the 1990s!

By the late 1990s, physiologically-motivated and perceptually-motivated approaches to signal processing began to flourish

Some major new trends ….

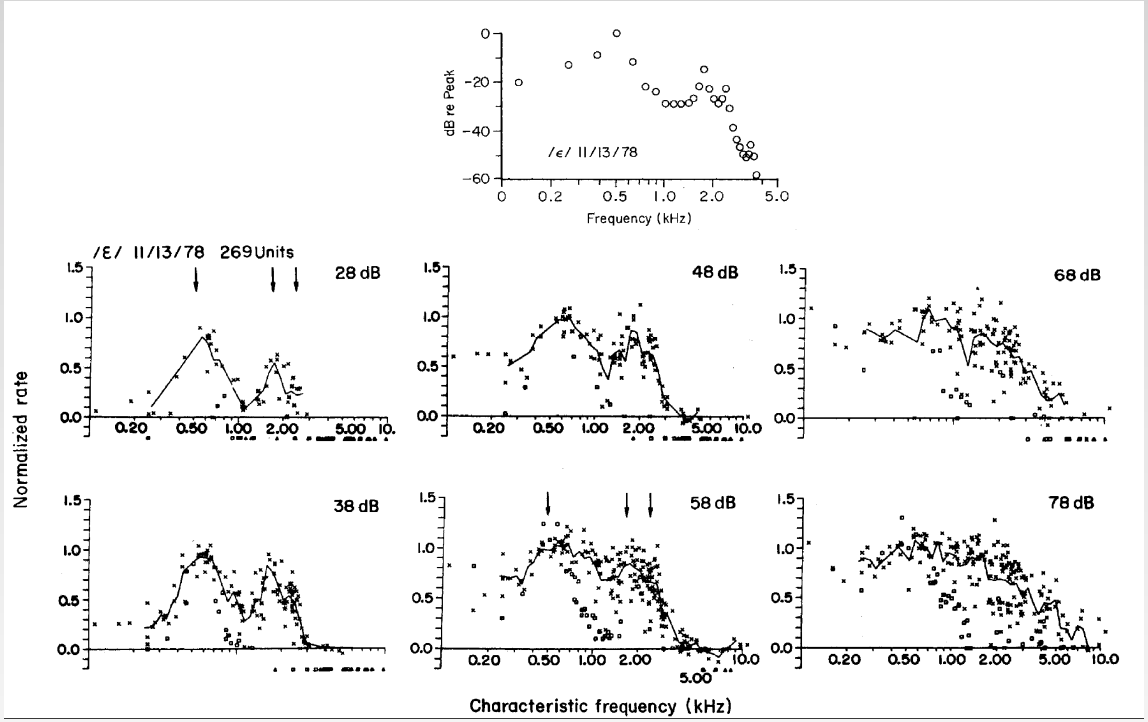
- **Computation no longer such a limiting factor**
- **Serious attention to temporal evolution**
- **Attention to reverberation**
- **Binaural processing**
- **More effective and mature approaches to information fusion**

Peripheral auditory modeling at CMU 2004–now

- **Foci of activities:**
  - Representing synchrony
  - The shape of the rate-intensity function
  - Revisiting analysis duration
  - Revisiting frequency resolution
  - Onset enhancement
  - Modulation filtering
  - Binaural and “polyaural” techniques
  - Auditory scene analysis: common frequency modulation

Speech representation using mean rate

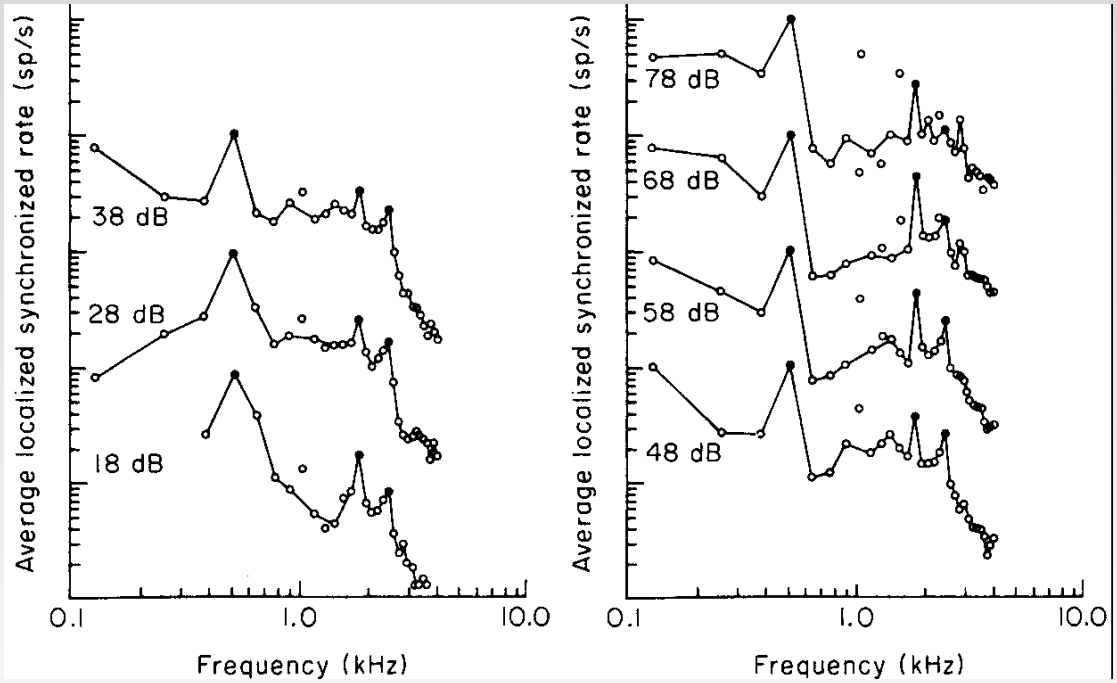
- Representation of vowels by Young and Sachs using mean rate:



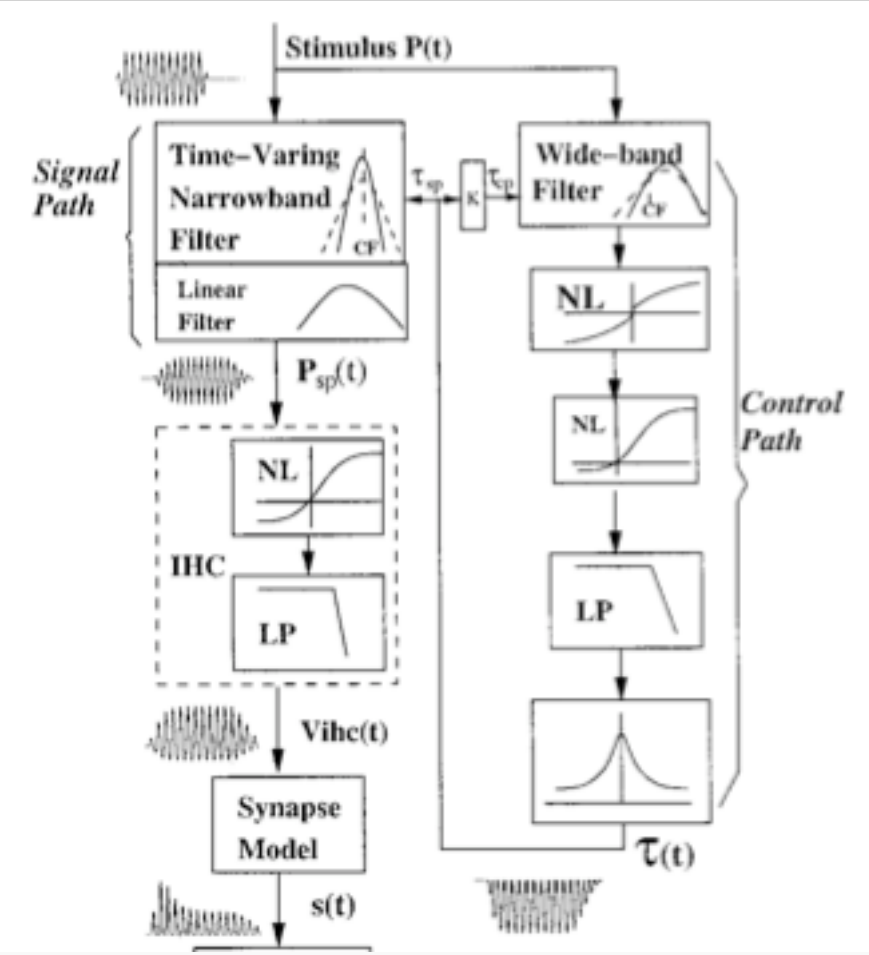
- Mean rate representation does not preserve spectral information

Speech representation using average localized synchrony rate

- **Representation of vowels by Young and Sachs using ALSR:**

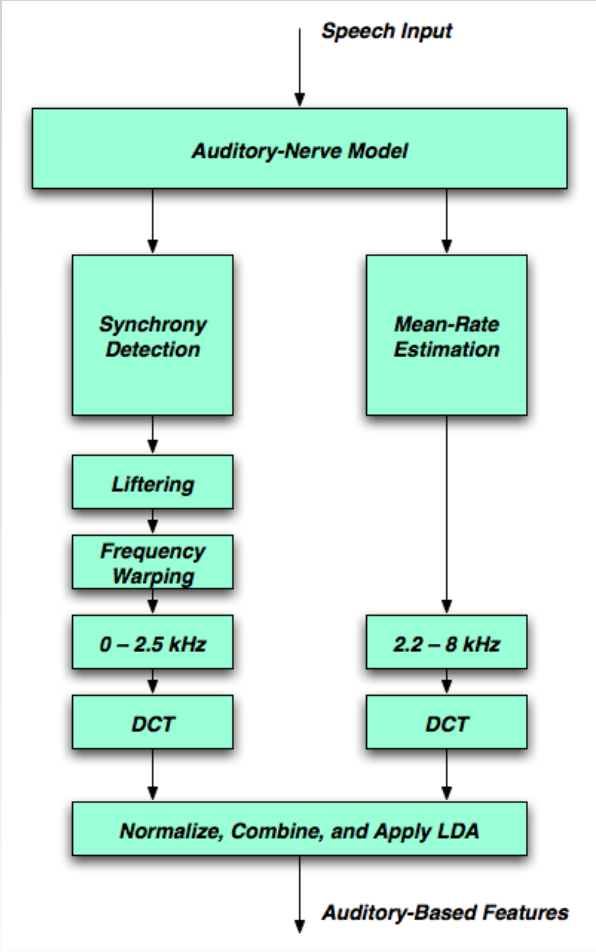


Physiologically-motivated signal processing: the Zhang-Carney model of the periphery



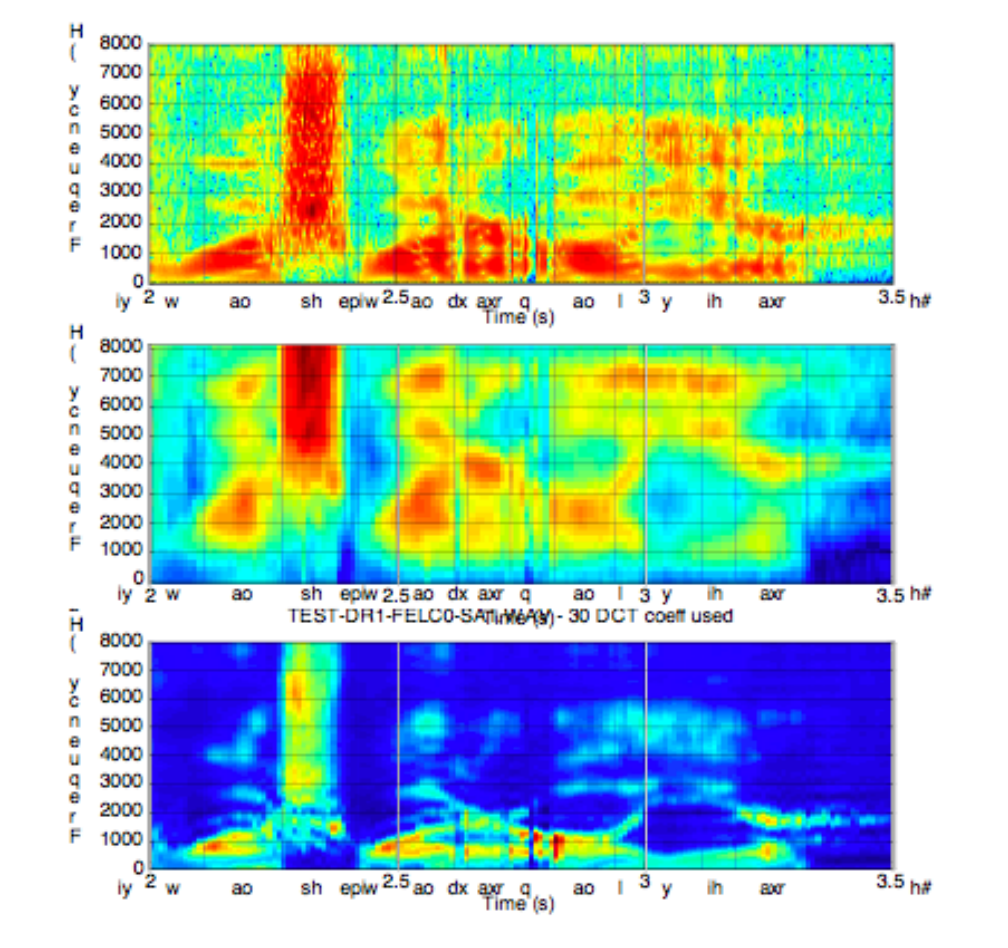
- We used the "synapse output" as the basis for further processing

Physiologically-motivated signal processing: synchrony and mean-rate detection (Kim/Chiu '06)



- 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

Comparing auditory processing with cepstral analysis: clean speech

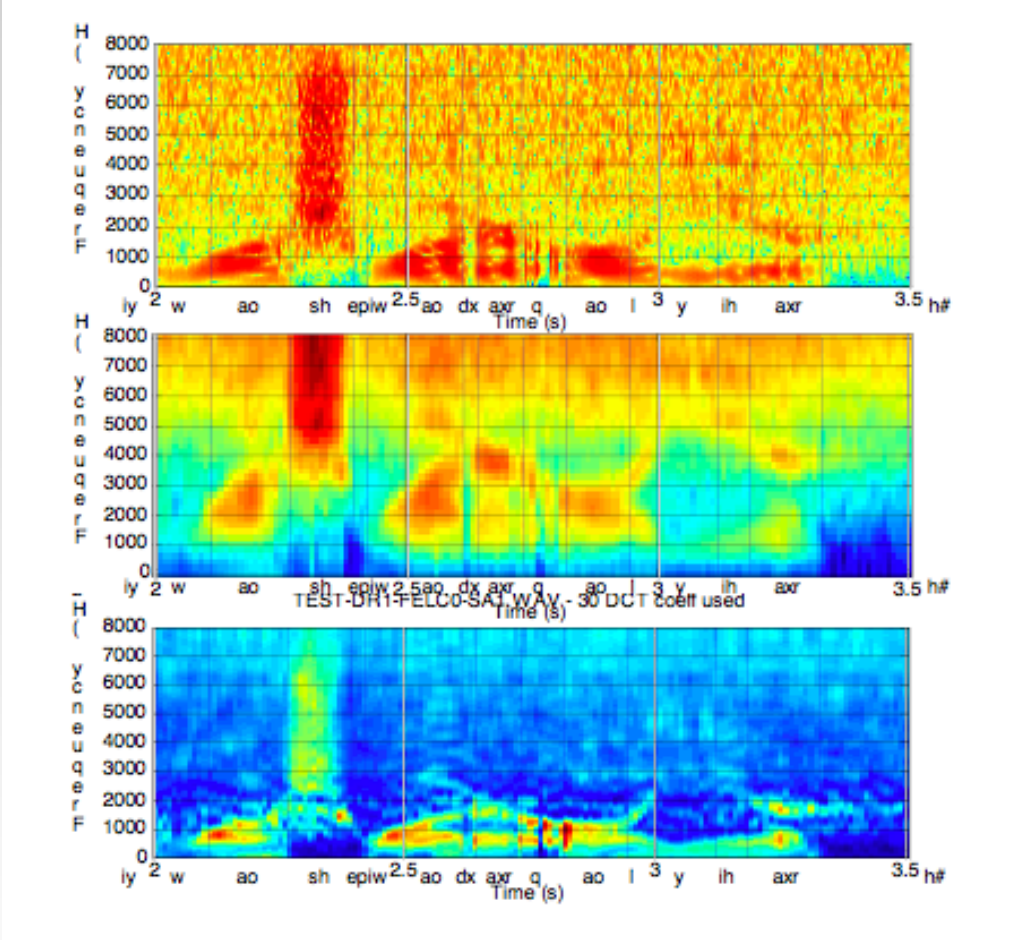


Original spectrogram

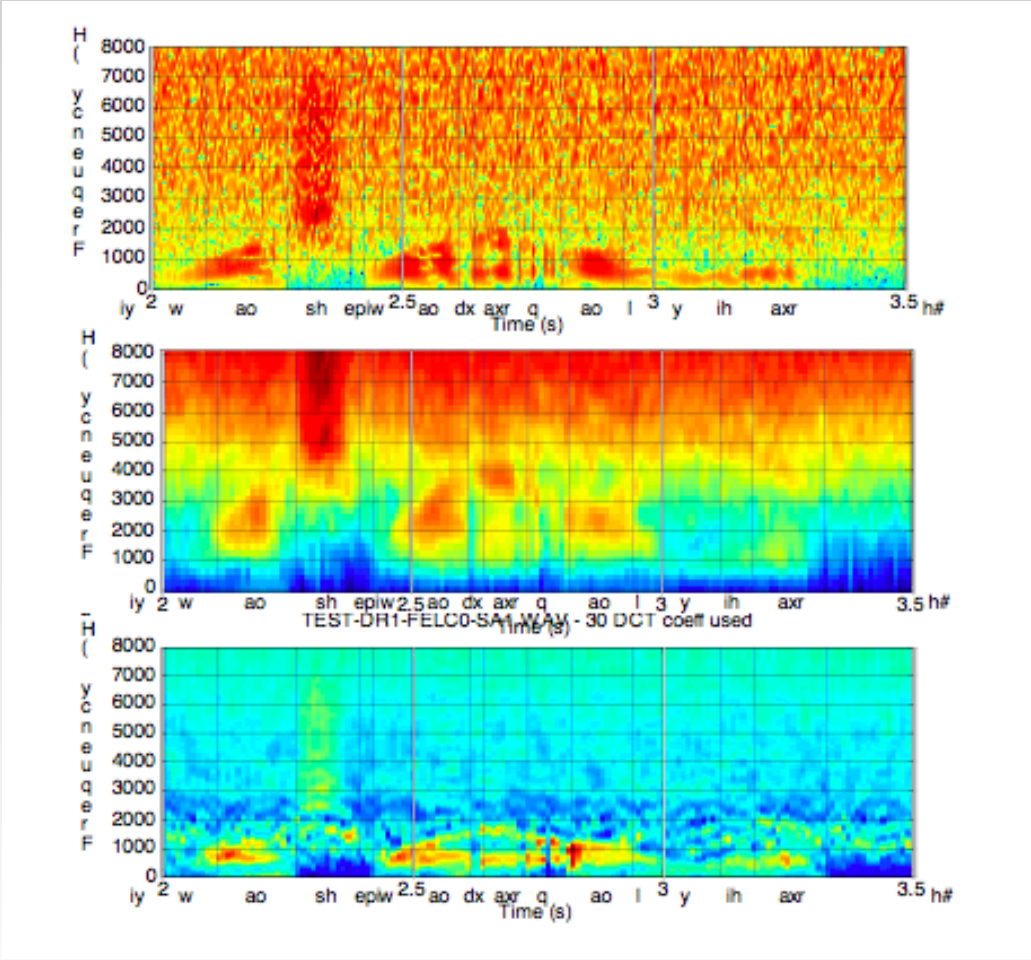
MFCC reconstruction

Auditory analysis

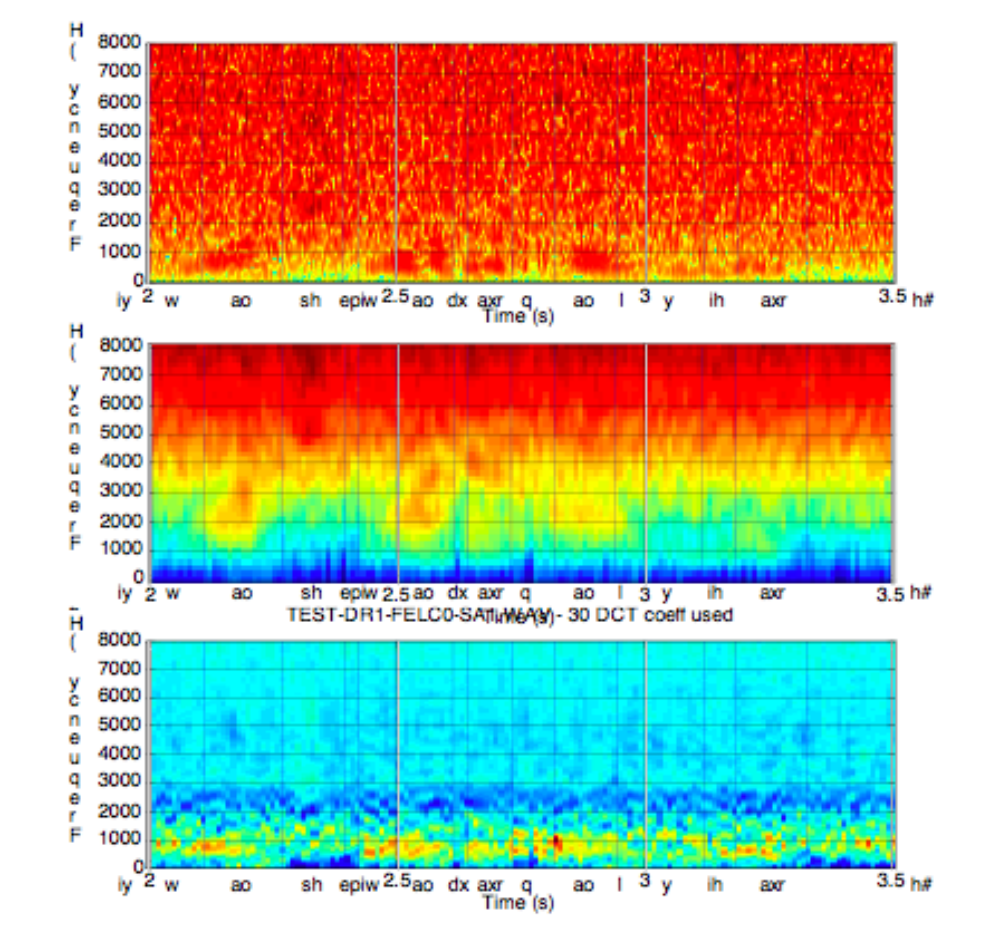
Comparing auditory processing with cepstral analysis: 20-dB SNR



Comparing auditory processing with cepstral analysis: 10-dB SNR

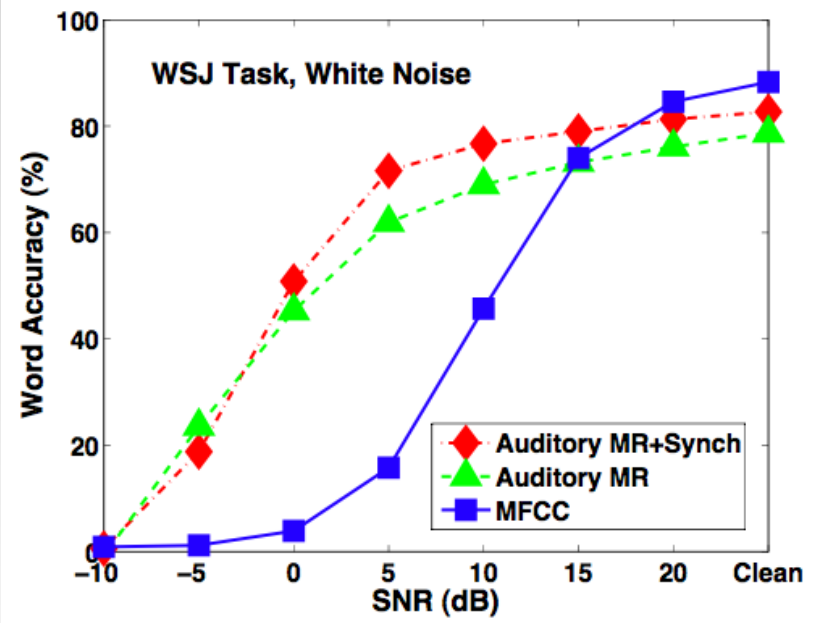


Comparing auditory processing with cepstral analysis: 0-dB SNR

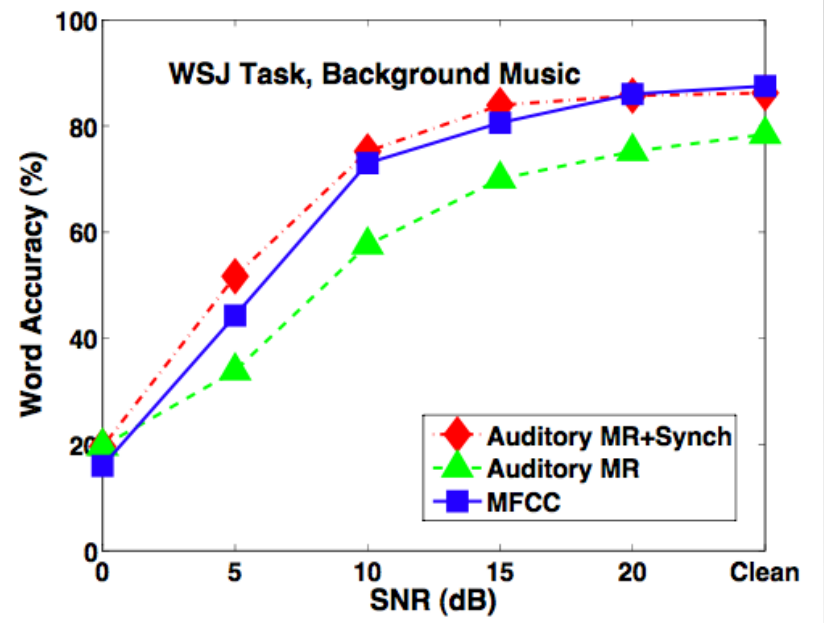


Auditory processing is more effective than MFCCs at low SNRs, especially in white noise

Accuracy in background noise:



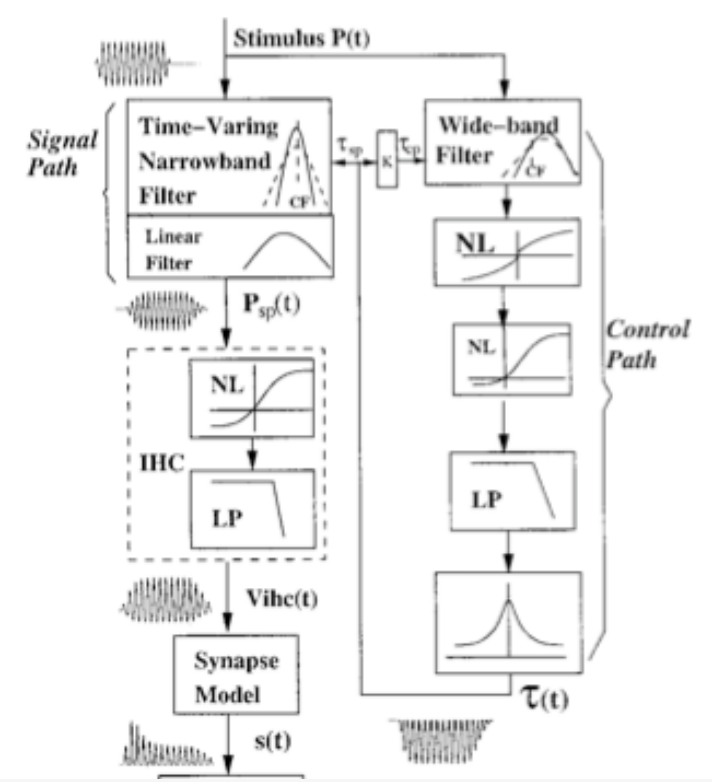
Accuracy in background music:



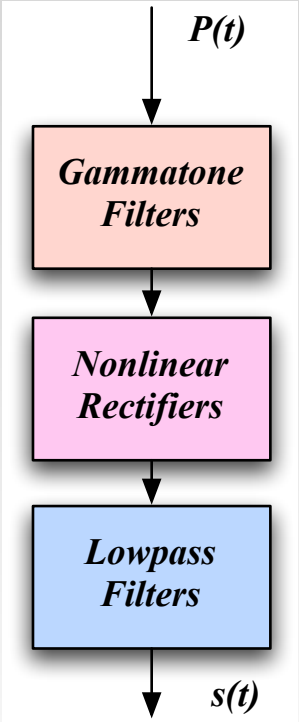
[Results from Kim et al., Interspeech 2006]

Do auditory models really need to be so complex?

- **Model of Zhang et al. 2001:**

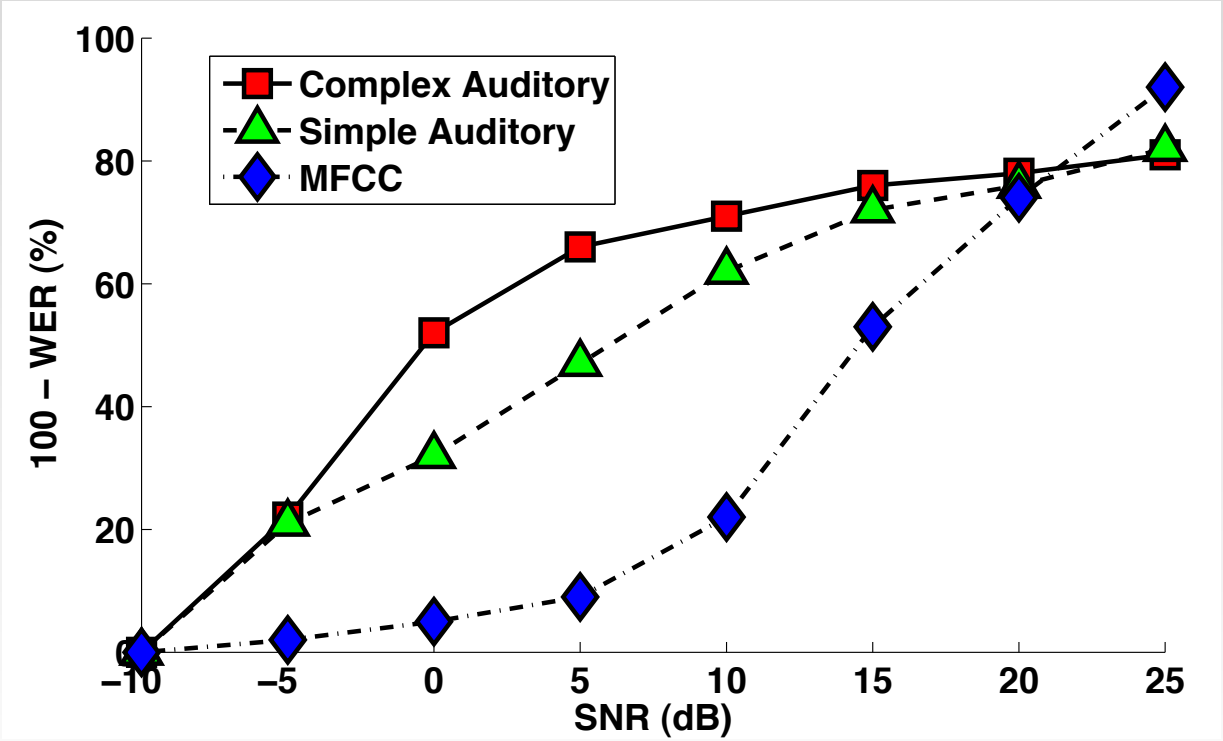


A much simpler model:

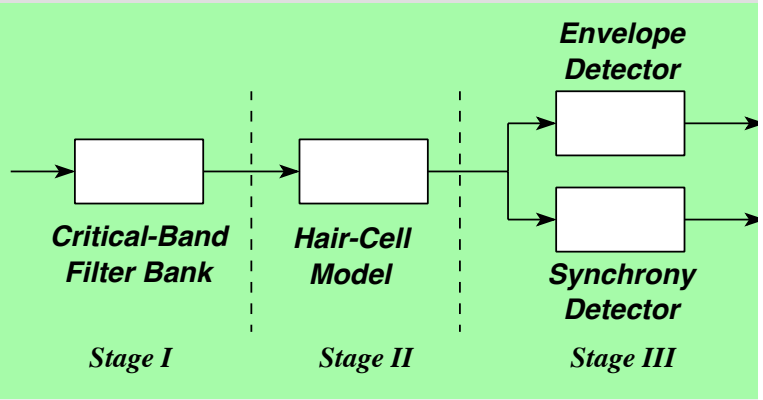


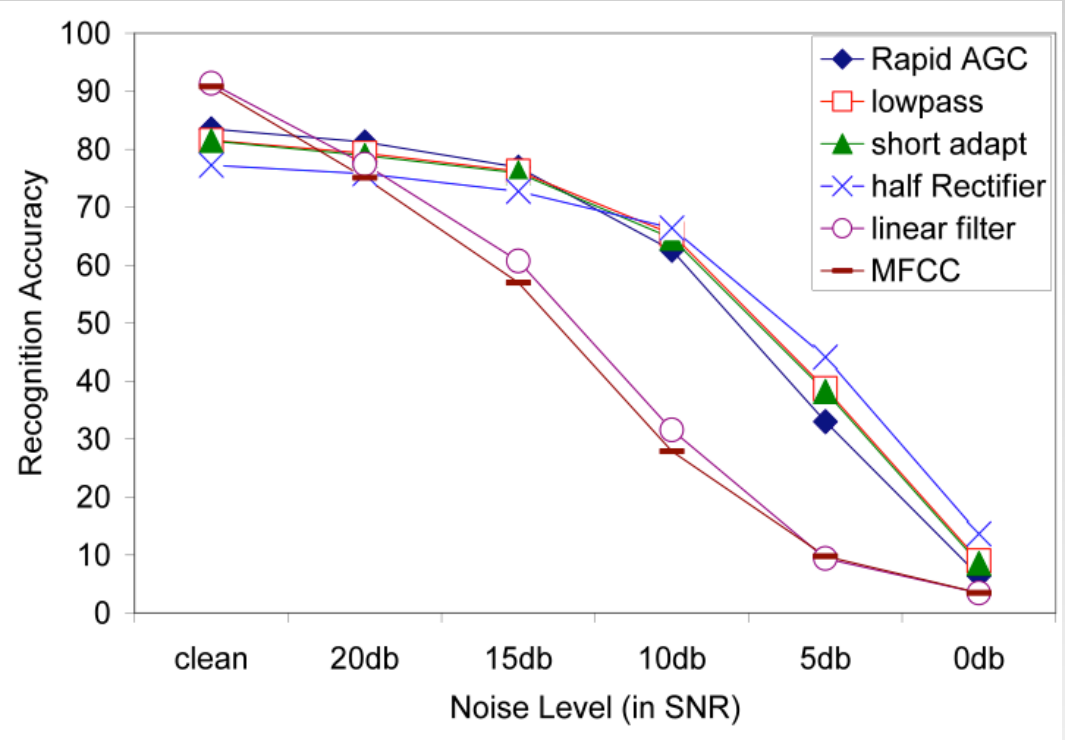
Comparing simple and complex auditory models

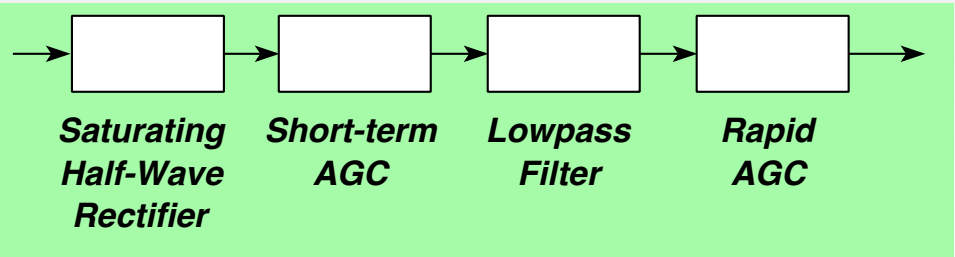
- Comparing MFCC processing, a trivial (filter–rectify–compress) auditory model, and the full Carney-Zhang model (Chiu 2006):



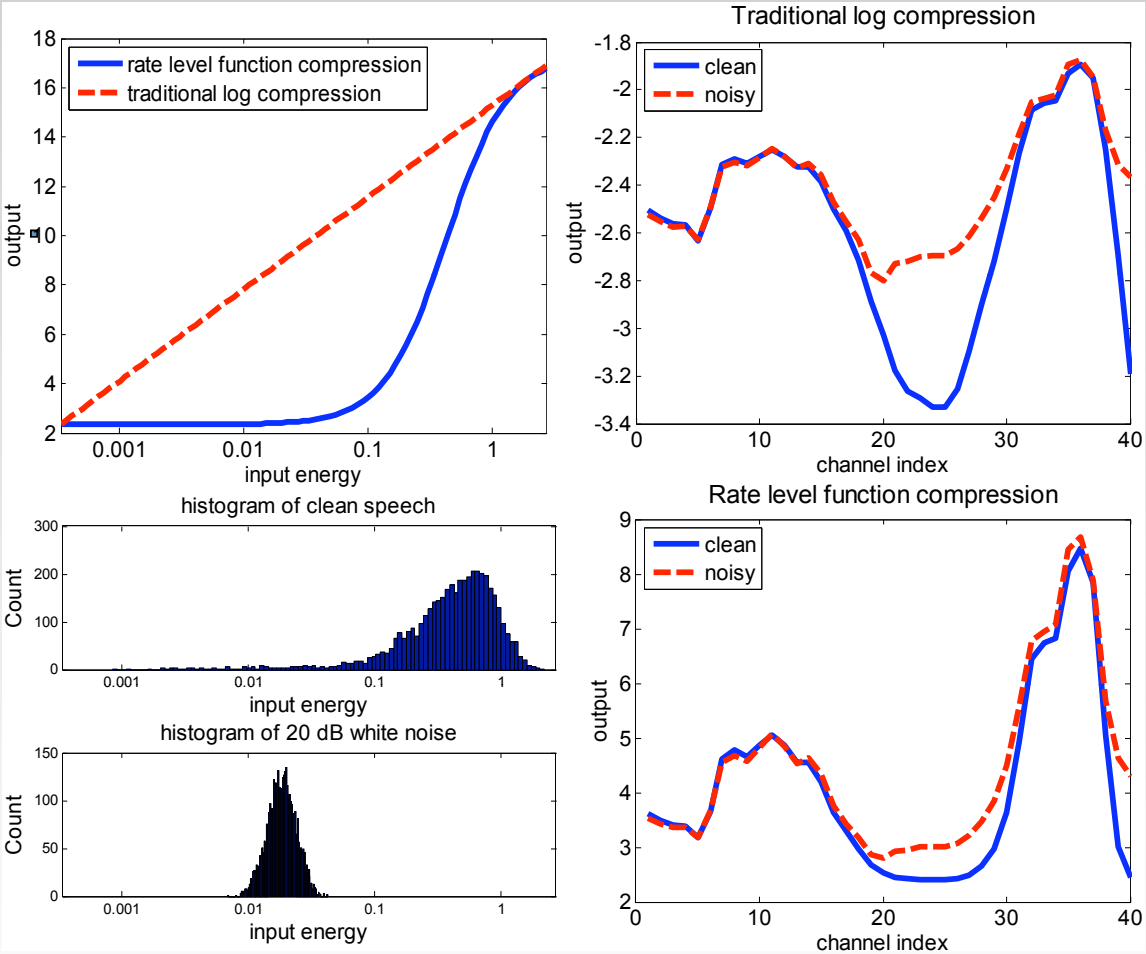
The nonlinearity seems to be the most important attribute of the Seneff model (Chiu '08)



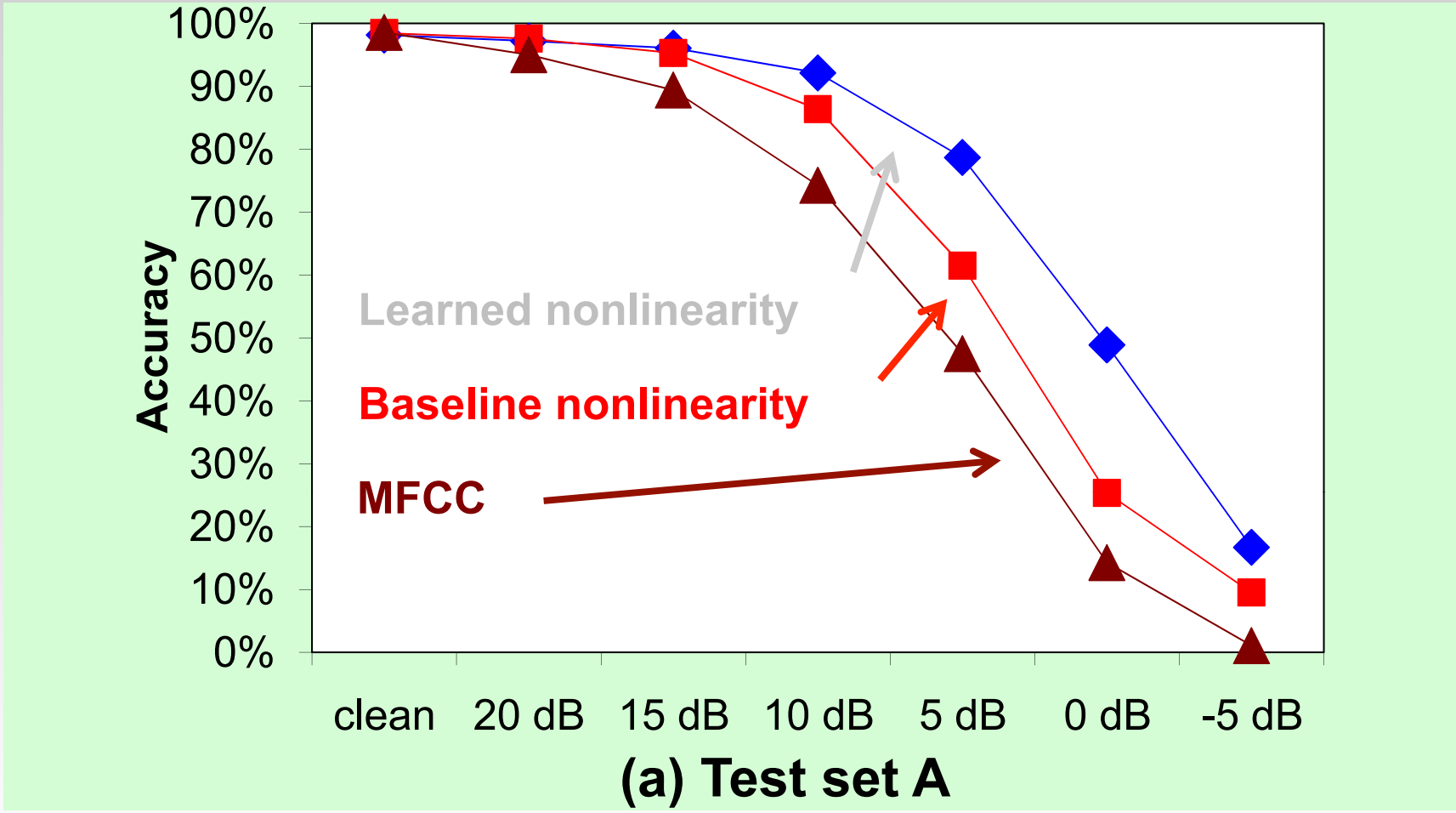




Why the nonlinearity seems to help …



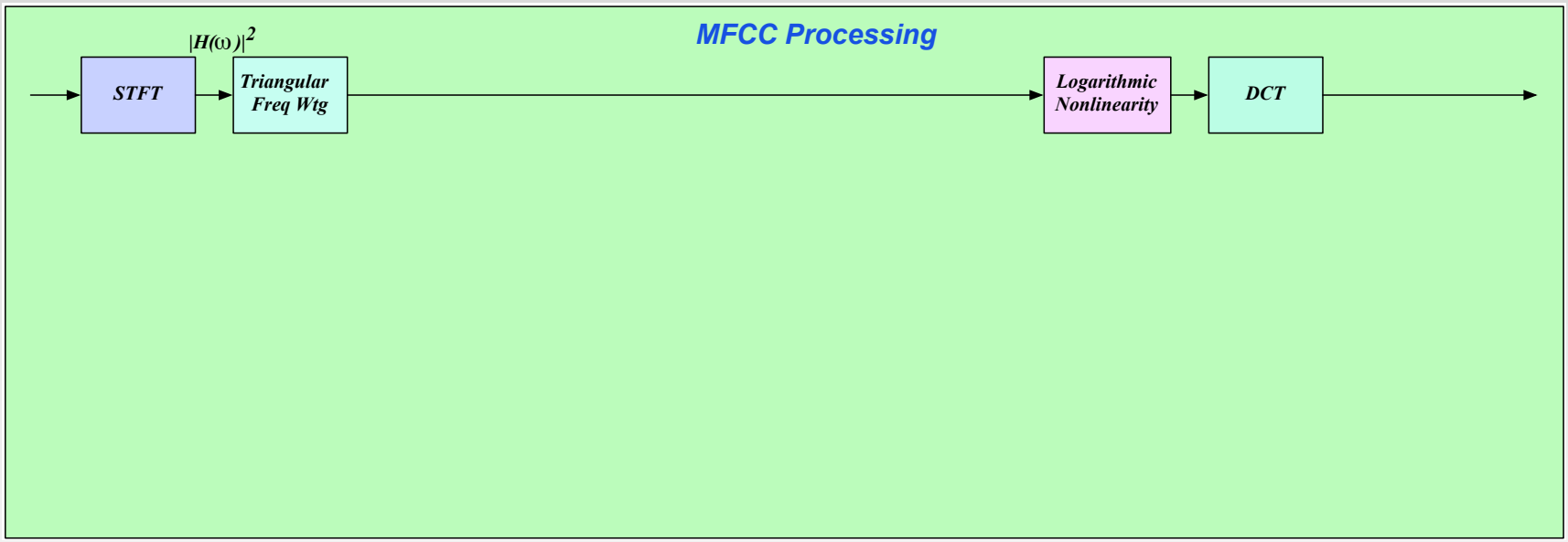
Impact of auditory nonlinearity (Chiu)



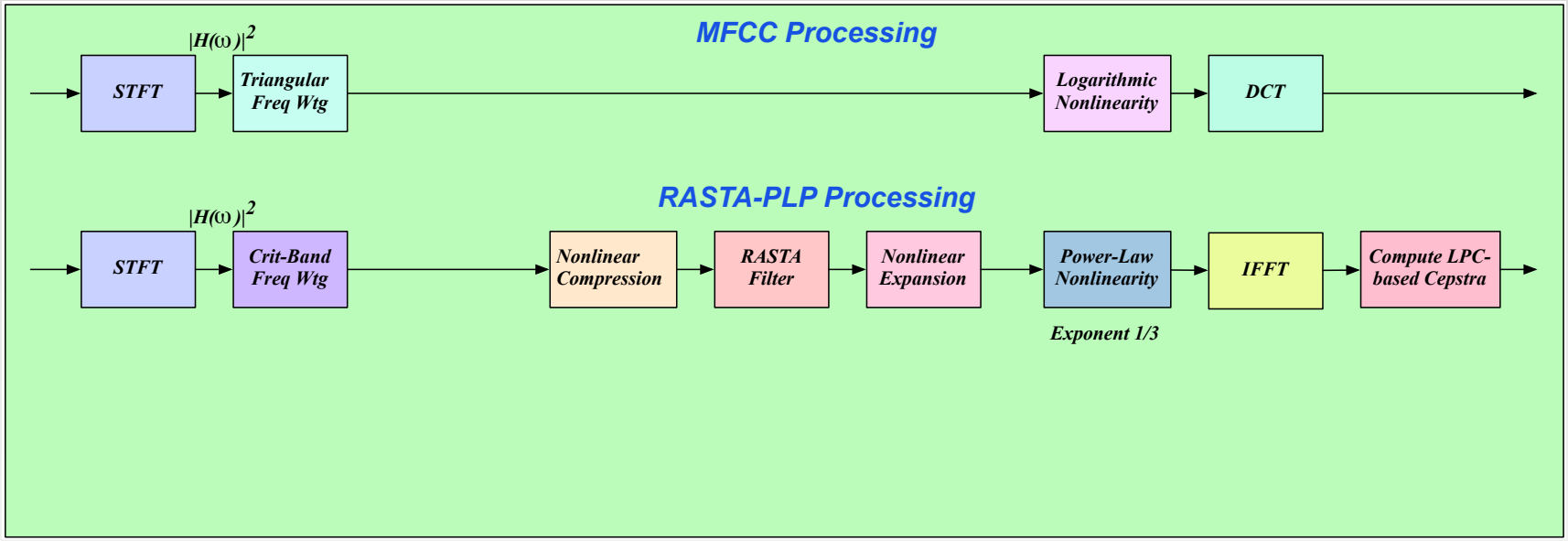
PNCC processing (Kim and Stern, 2010,2014)

- **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

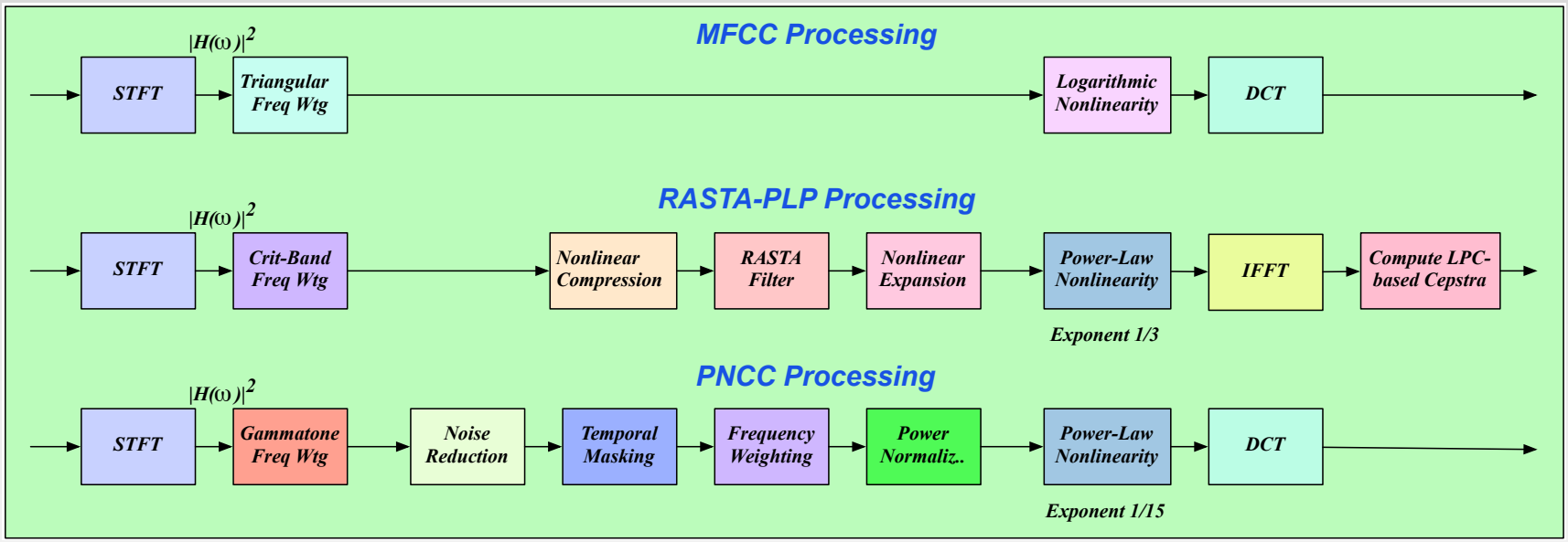
An integrated front end: power-normalized cepstral coefficients (PNCC, Kim '10)



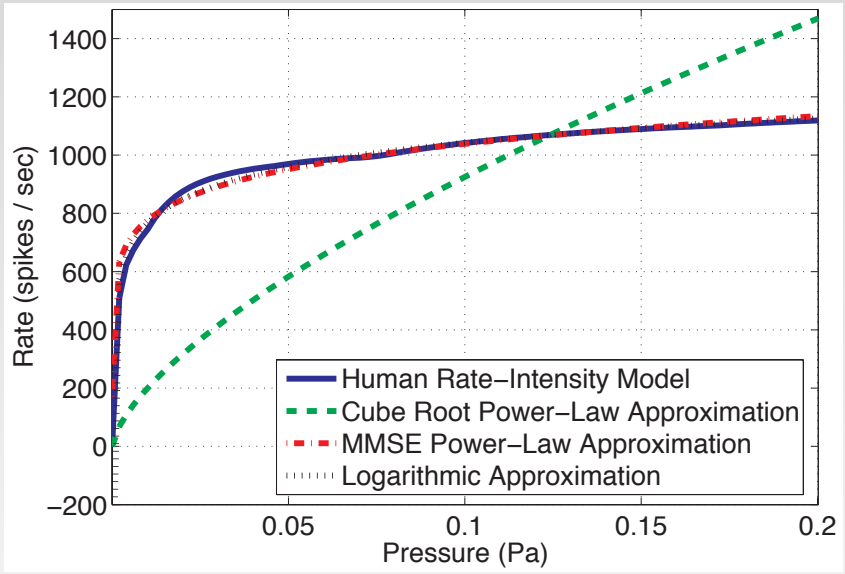
An integrated front end: power-normalized cepstral coefficients (PNCC, Kim '10)

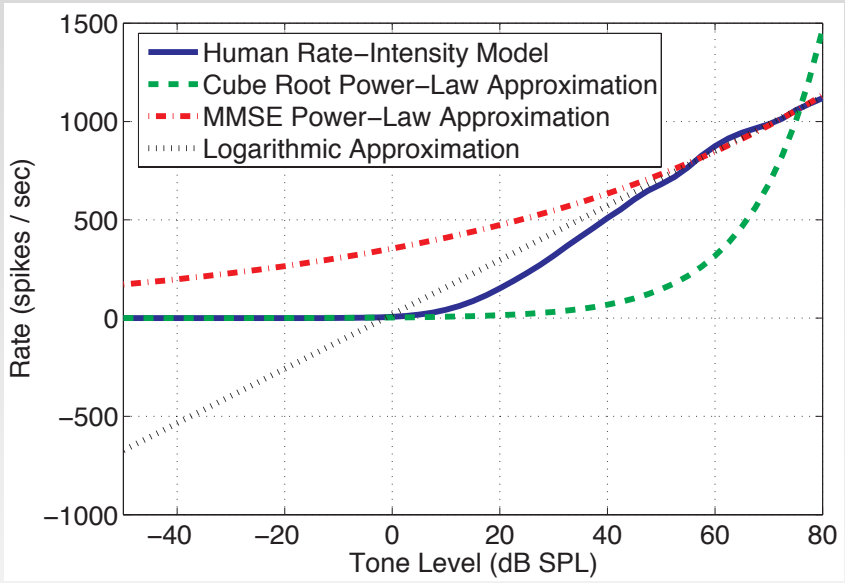


An integrated front end: power-normalized cepstral coefficients (PNCC, Kim '10)



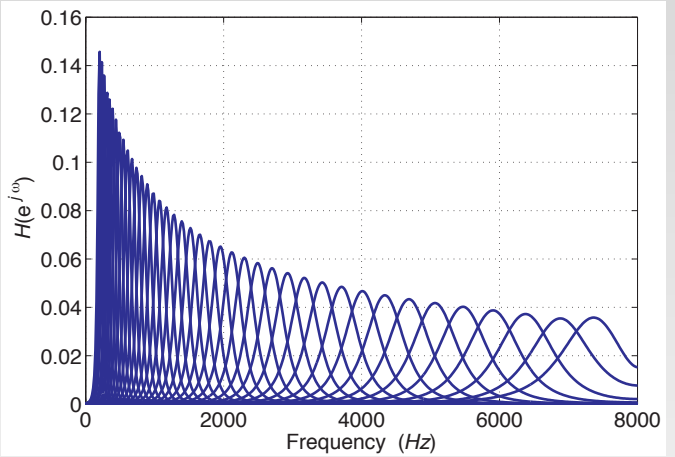
The nonlinearity in PNCC processing (Kim)





Frequency resolution

- **Examined several types of frequency resolution**
  - MFCC triangular filters
  - Gammatone filter shapes
  - Truncated Gammatone filter shapes



- **Most results do not depend greatly on filter shape**
- **Some sort of frequency integration is helpful when frequency-based selection algorithms are used**

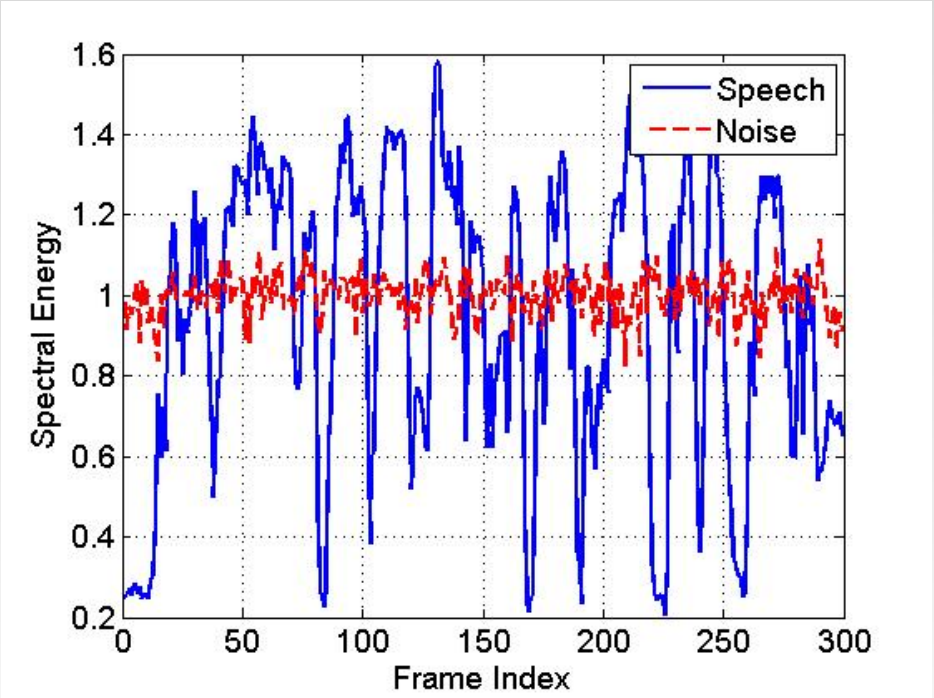
Analysis window duration (Kim)

- Typical analysis window duration for speech recognition is ~25-35 ms
- Optimal analysis window duration for estimation of environmental parameters is ~75-120 ms
- Best systems measure environmental parameters (including voice activity detection over a longer time interval but apply results to a short-duration analysis frame

Temporal Speech Properties: modulation filtering

Output of speech and noise segments from 14th Mel filter (1050 Hz)

- Speech segment exhibits greater fluctuations



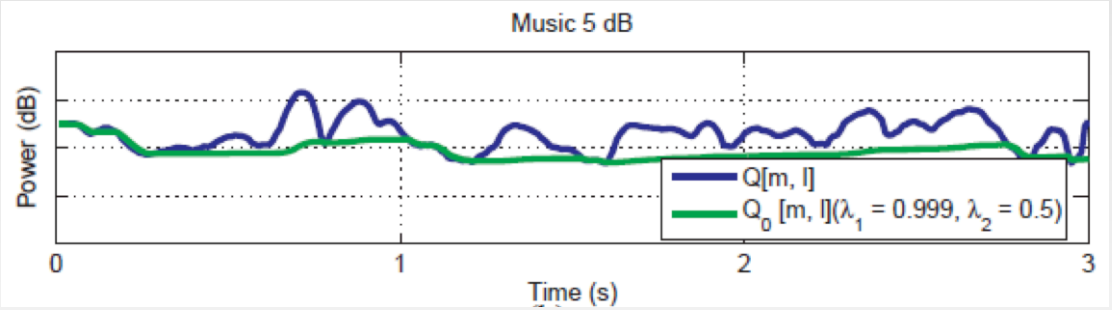
Nonlinear noise processing

- **Use nonlinear cepstral highpass filtering to pass speech but not noise**

- **Why nonlinear?**
  - Need to keep results positive because we are dealing with manipulations of signal power

Asymmetric lowpass filtering (Kim, 2010)

- **Overview of processing:**
  - Assume that noise components vary slowly compared to speech components
  - Obtain a running estimate of noise level in each channel using nonlinear processing
  - Subtract estimated noise level from speech
- **An example:**



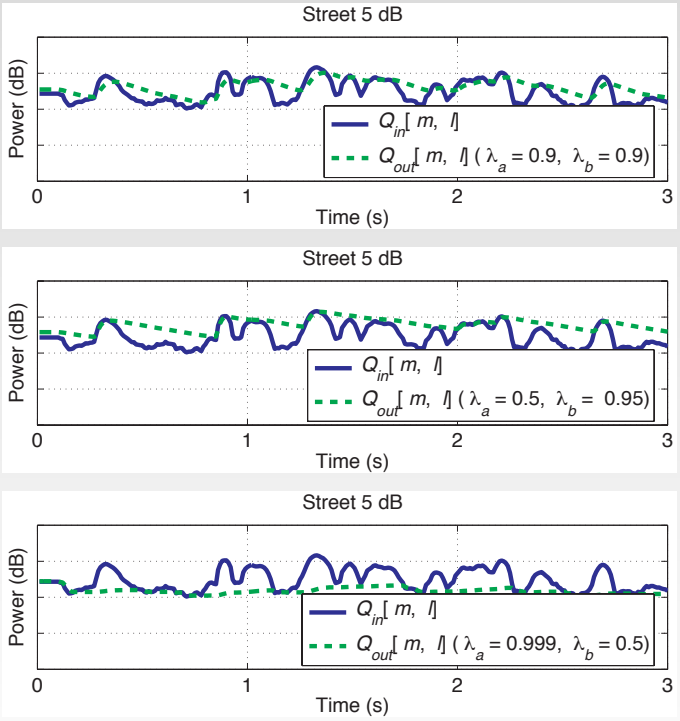
Note: Asymmetric highpass filtering is obtained by subtracting the lowpass filter output from the input

Implementing asymmetric lowpass filtering

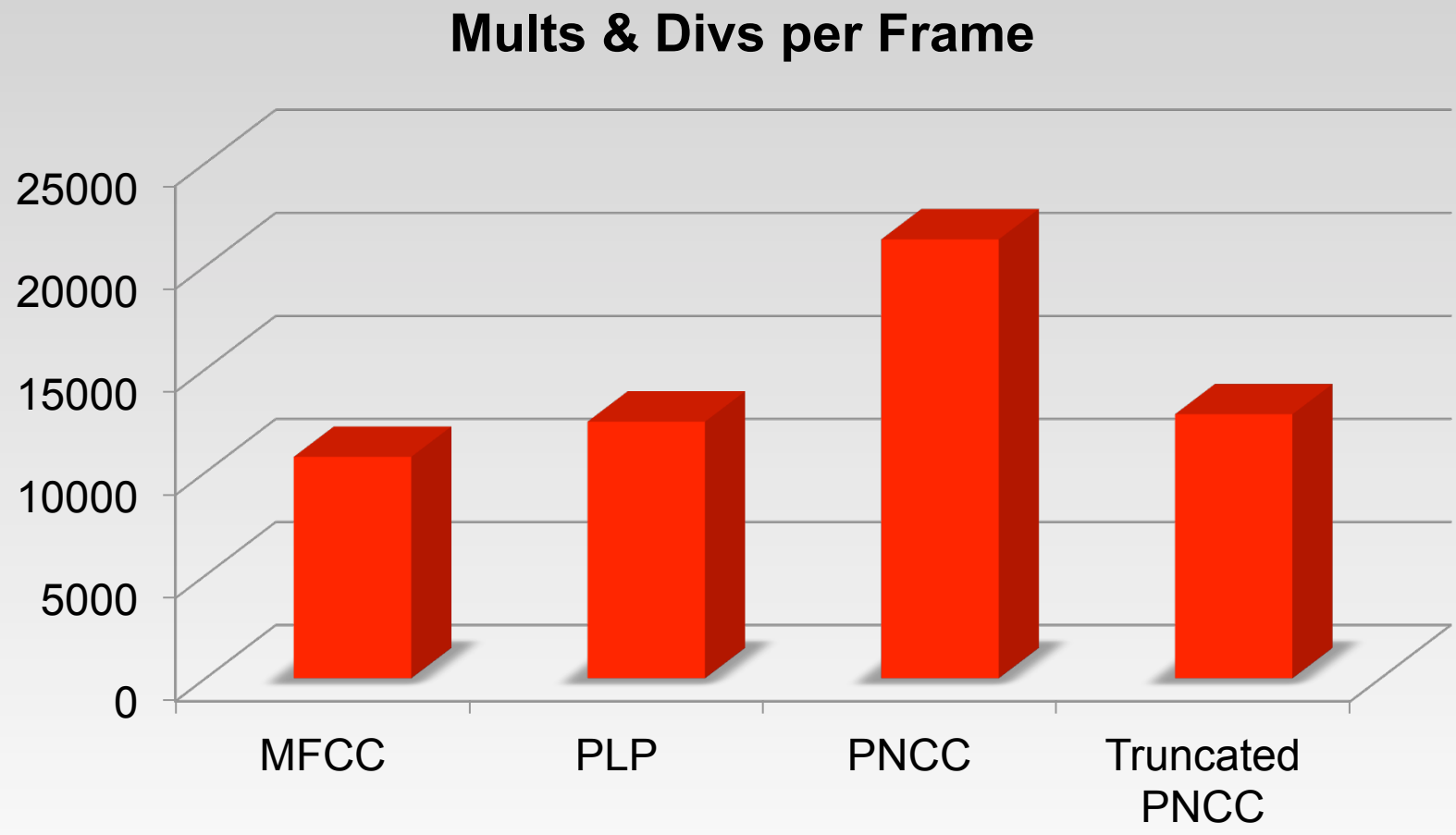
Basic equation:

$$Q_{out}[m,l] = \begin{cases} \lambda_a Q_{out}[m-1,l] + (1-\lambda_a) Q_{in}[m,l], \\ \quad \text{if } Q_{in}[m,l] \geq Q_{out}[m,l] \\ \lambda_b Q_{out}[m-1,l] + (1-\lambda_b) Q_{in}[m,l], \\ \quad \text{if } Q_{in}[m,l] < Q_{out}[m,l] \end{cases}$$

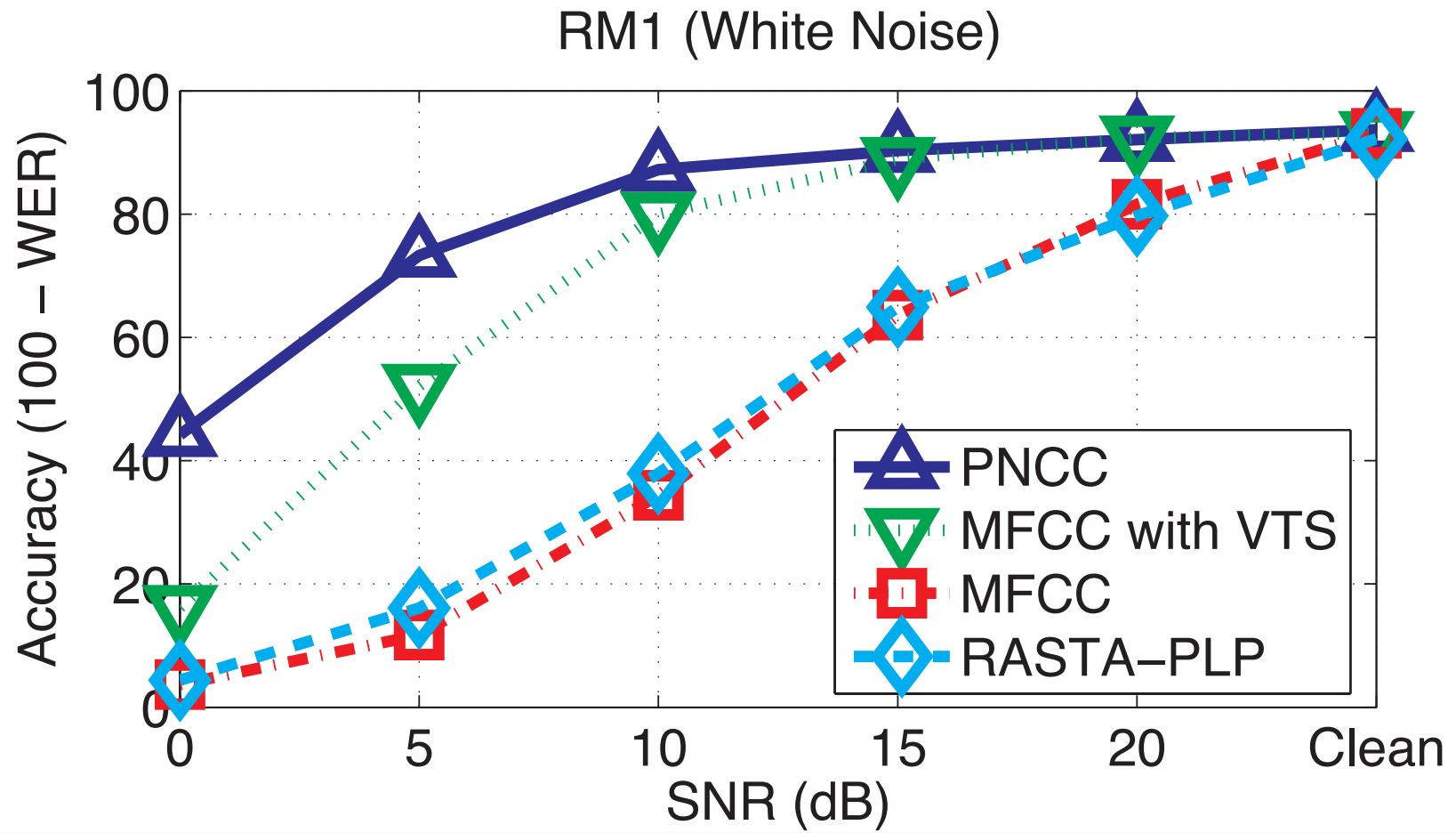
Dependence on parameter values:



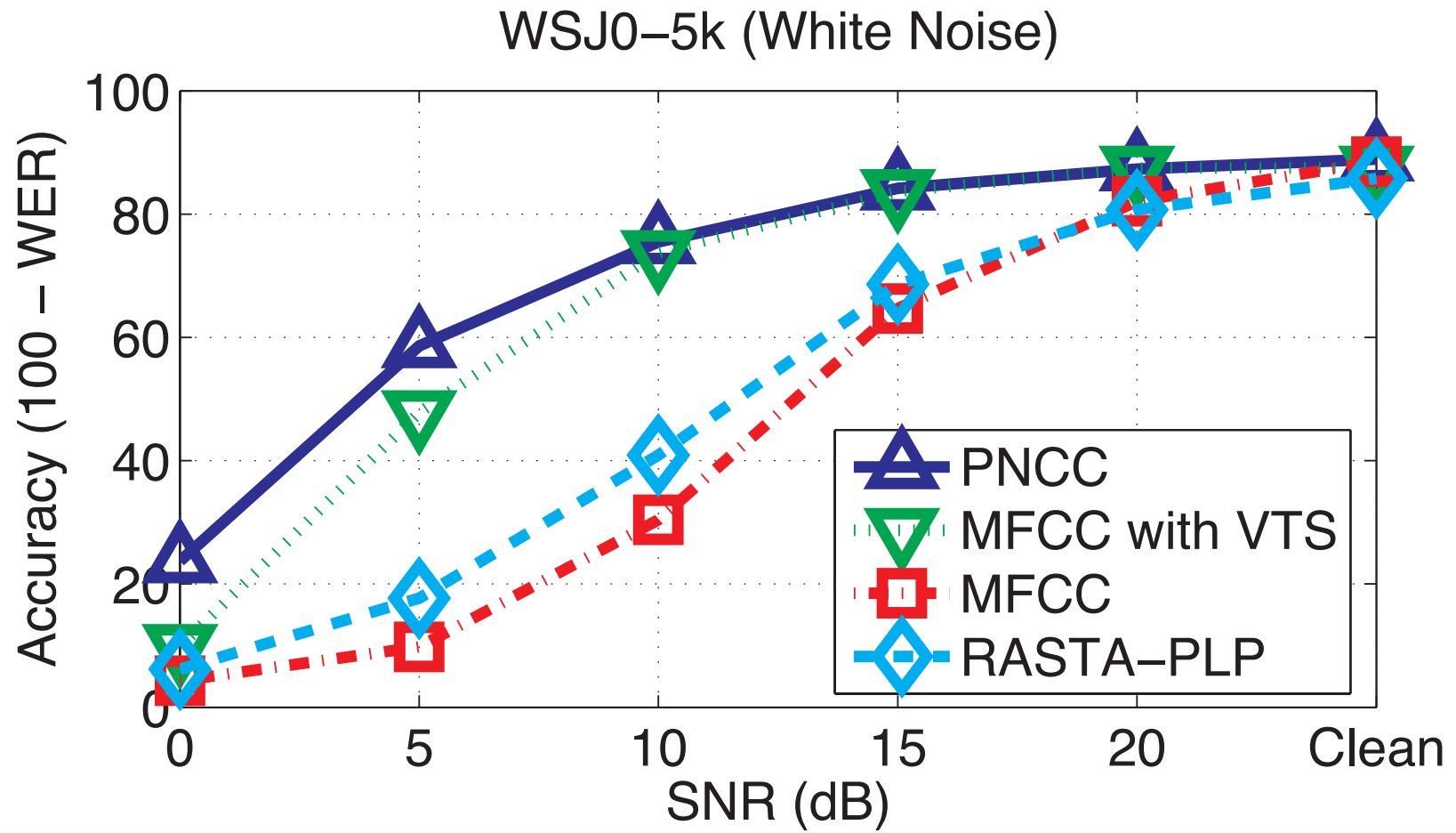
Computational complexity of front ends



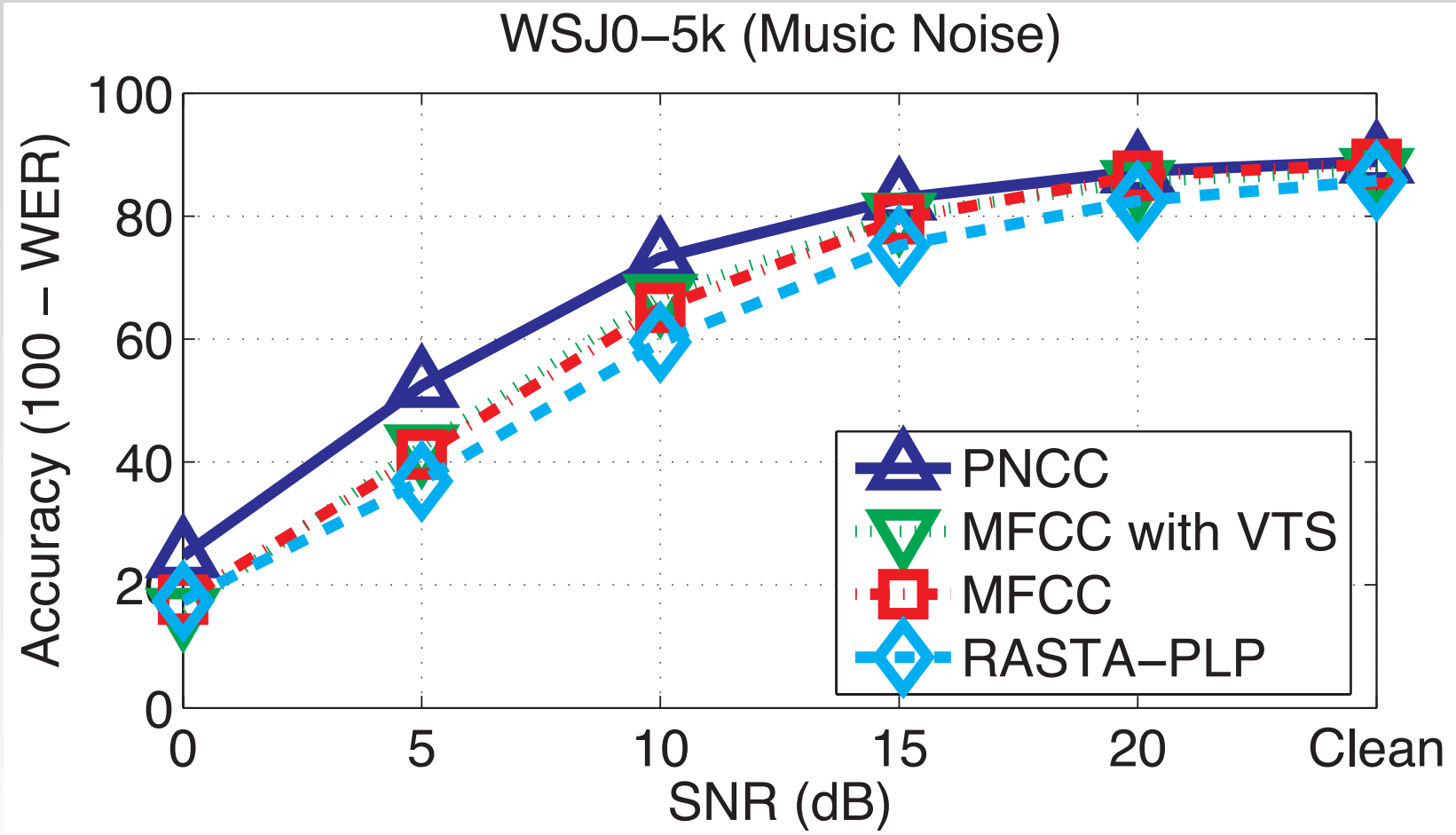
Performance of PNCC in white noise (RM)



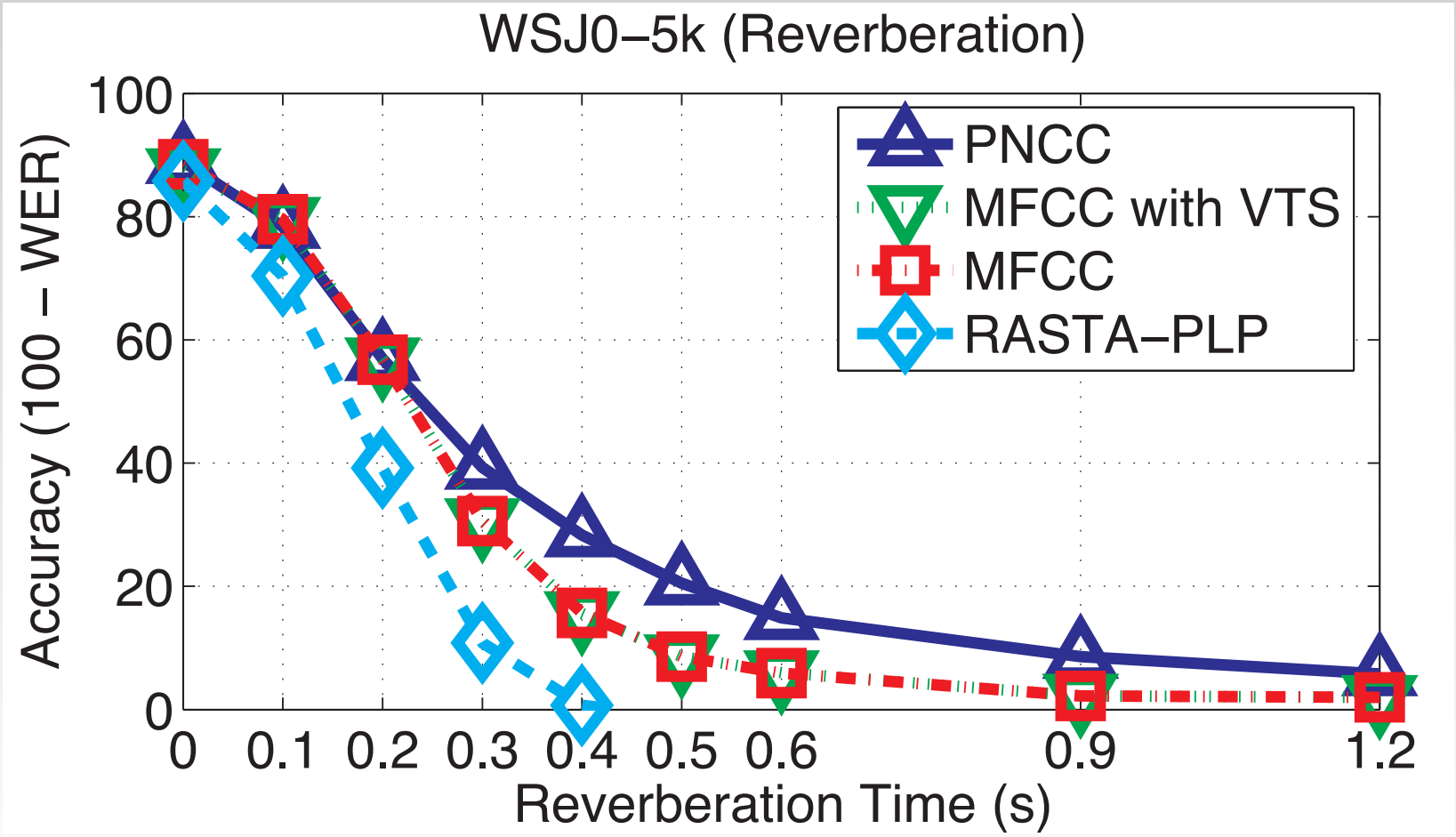
Performance of PNCC in white noise (WSJ)



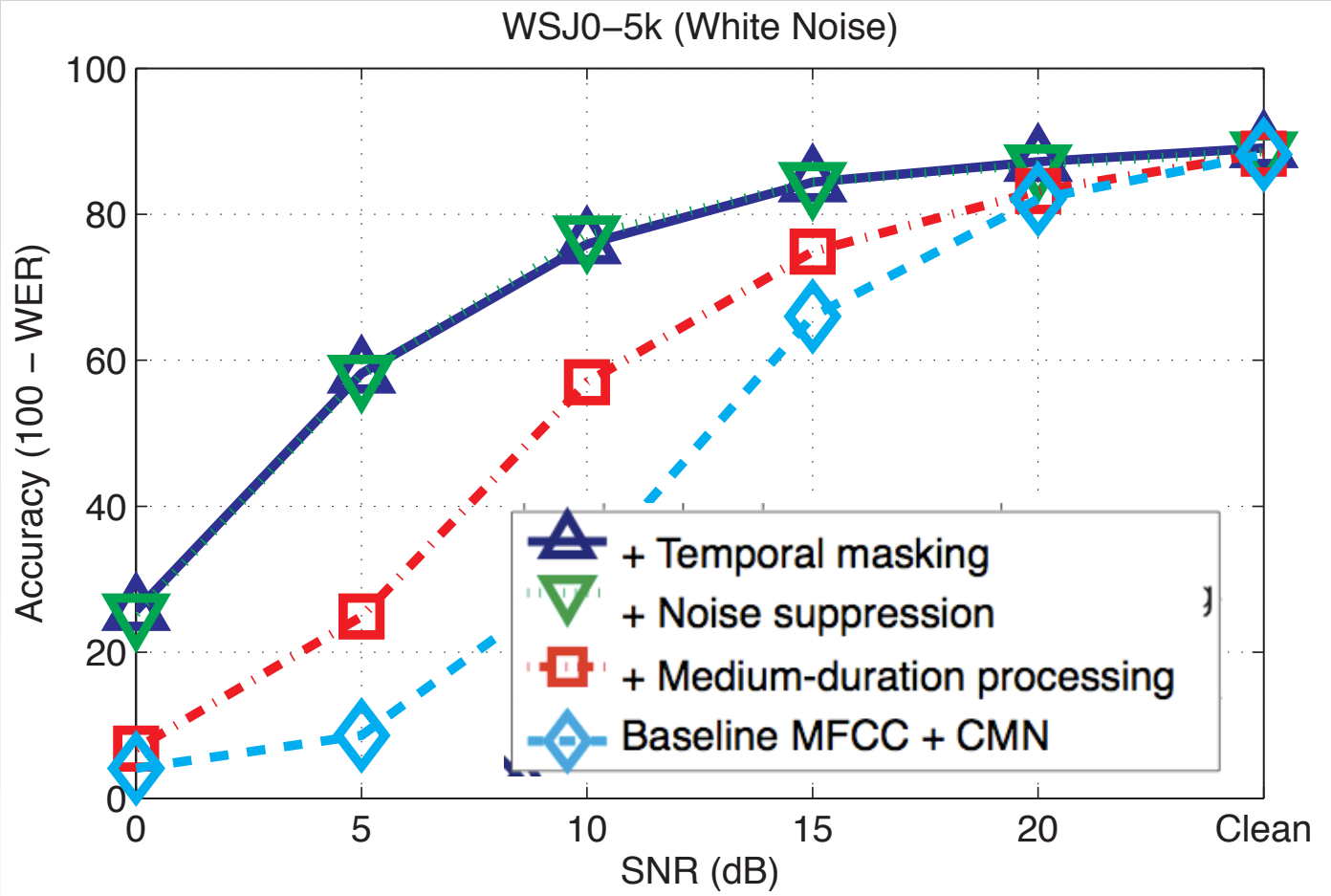
Performance of PNCC in background music



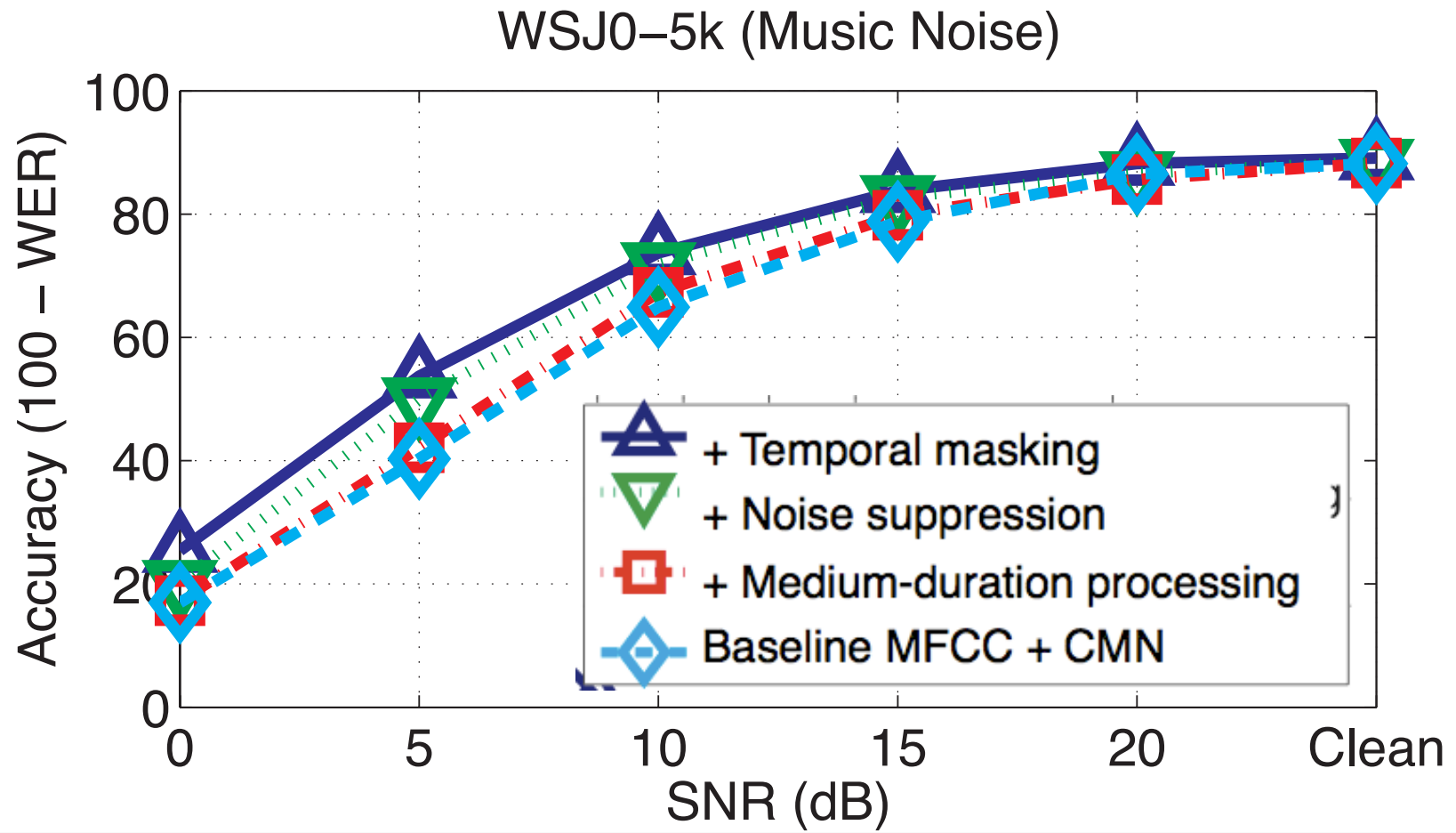
Performance of PNCC in reverberation



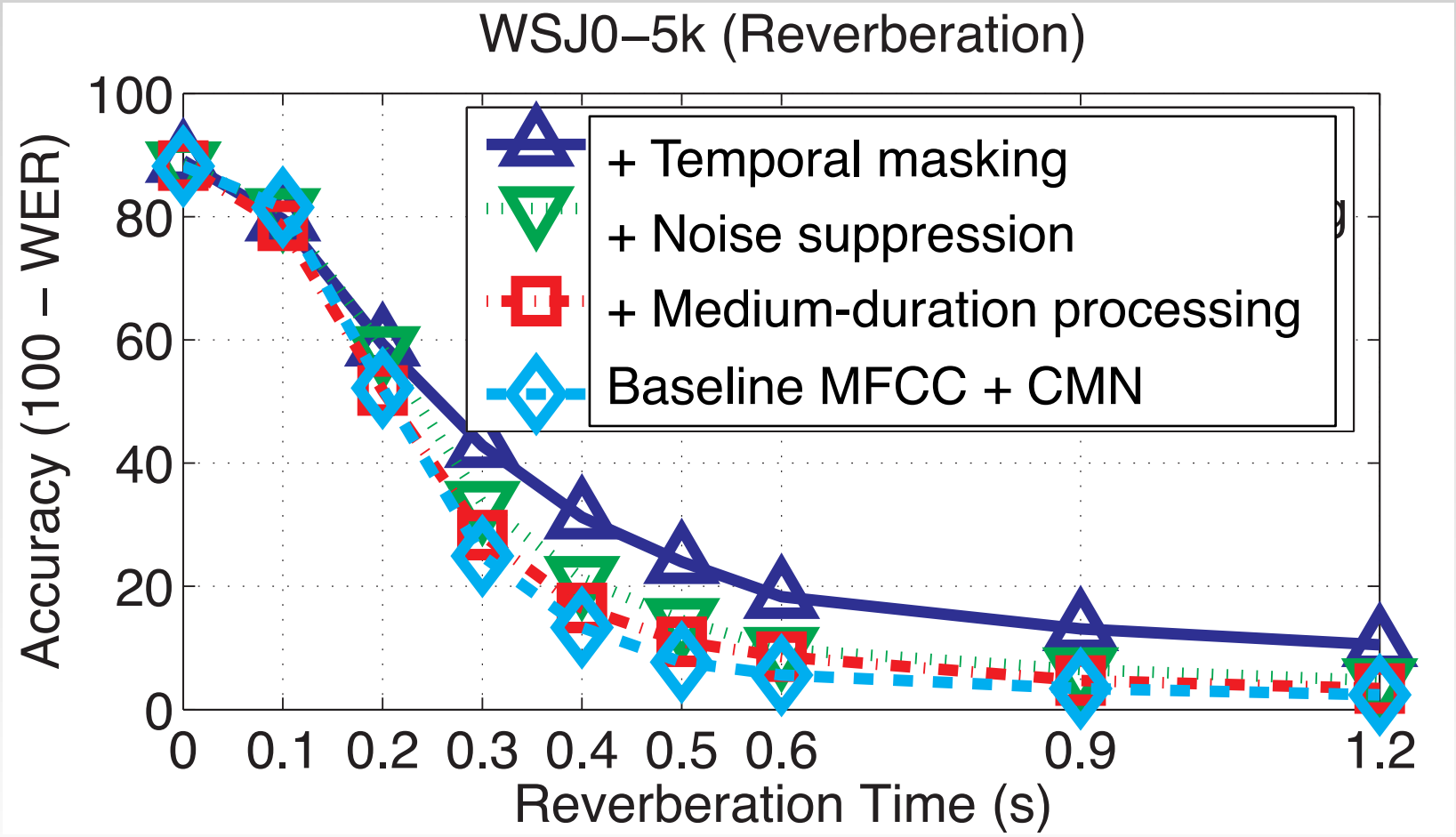
Contributions of PNCC components: white noise (WSJ)



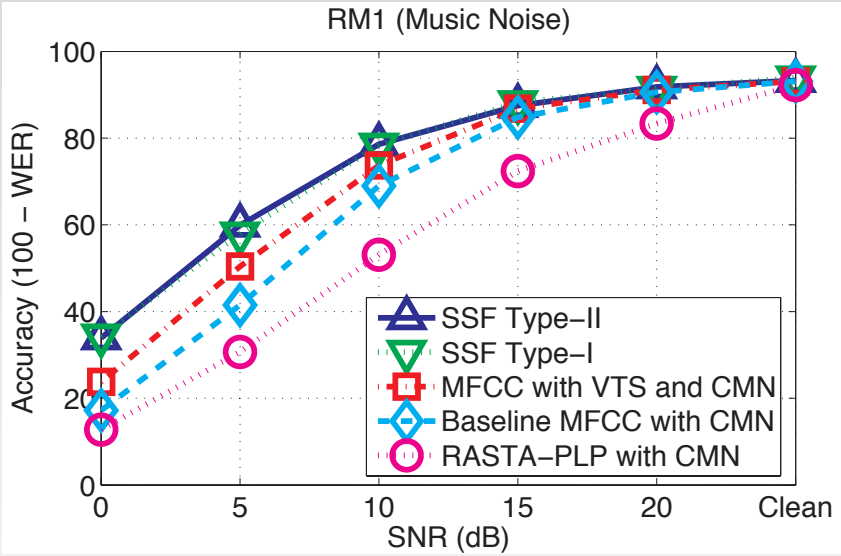
Contributions of PNCC components: background music (WSJ)

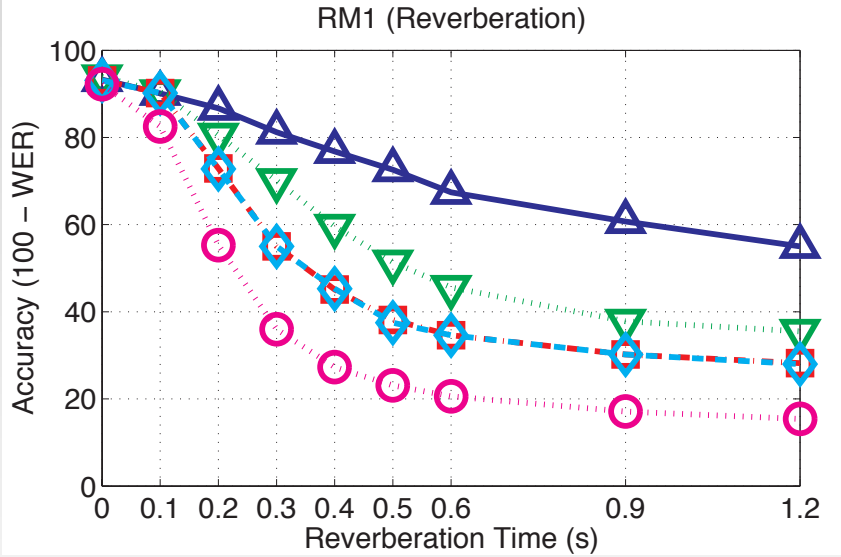


Contributions of PNCC components: reverberation (WSJ)

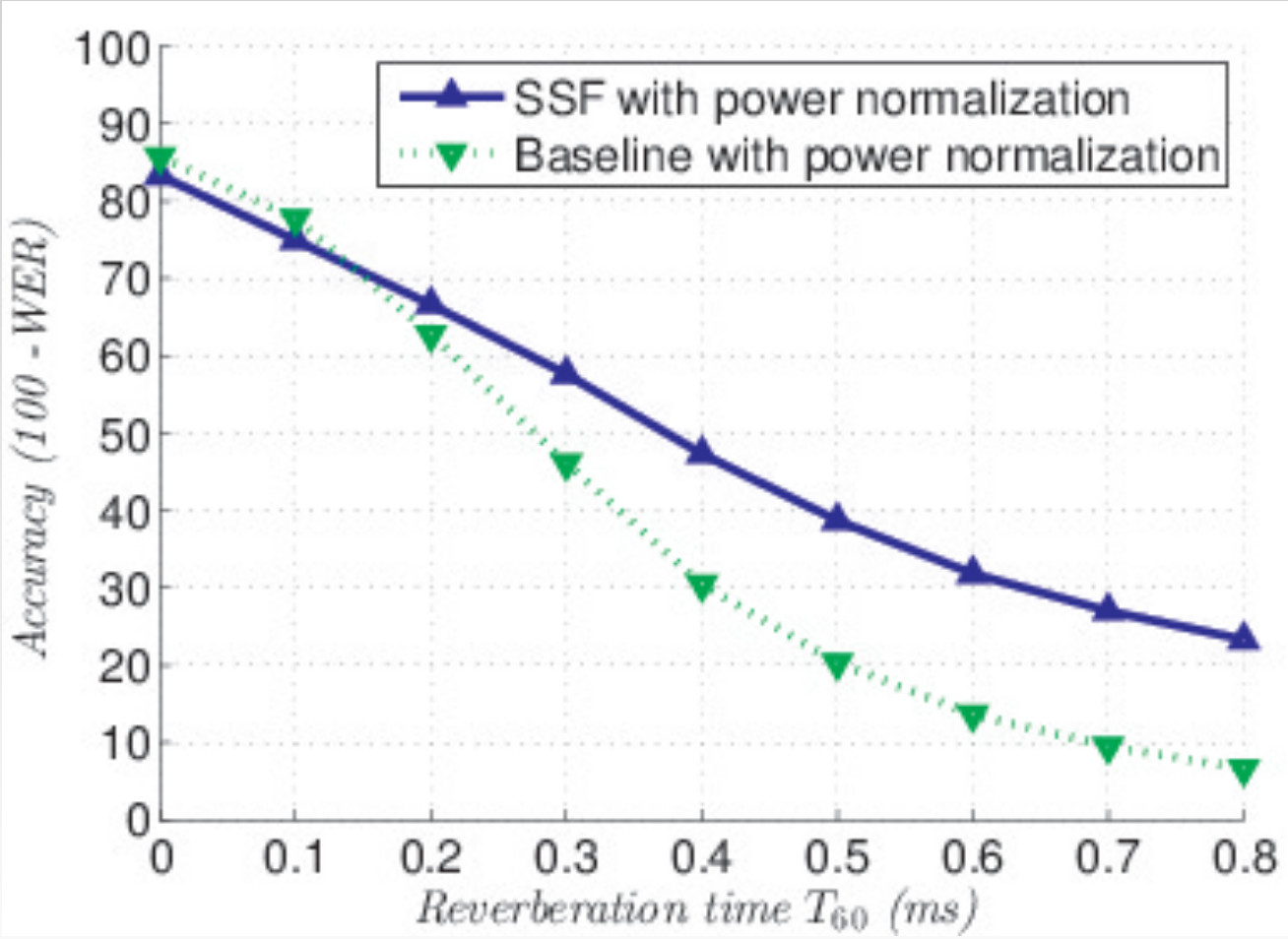


Effects of onset enhancement/temporal masking (SSF processing, Kim '10)





PNCC and SSF @Google



Summary … so what matters?

- **Knowledge of the auditory system can certainly improve ASR accuracy:**
  - Use of synchrony
  - Consideration of rate-intensity function
  - Onset enhancement
  - Nonlinear modulation filtering
  - Selective reconstruction
  - Consideration of processes mediating auditory scene analysis

Summary: PNCC processing

- **PNCC processing includes**
  - More effective nonlinearity
  - Parameter estimation for noise compensation and analysis based on longer analysis time and frequency spread
  - Efficient noise compensation based on modulation filtering
  - Onset enhancement
  - Computationally-efficient implementation
- **Not considered yet …**
  - Synchrony representation
  - Lateral suppression