Should Recognizers Have Ears? (Hermansky; 1998)

Pedro Davalos

CPSC 680-604

Feb 8, 2007

Outline

- Background
- Analysis
- Spectral Domain (PLP)
- Temporal Domain (RASTA)
- Partial Information
- Conclusions

Background: Traditional ASR

- Tradeoff between Knowledge and Training Data
  - Built-in knowledge would make a redundant ASR
  - Successful “Traditional” ASRs require:
    - Extensive training data for each particular application
    - Extremely controlled environment
- “Traditional” Statistical ASRs (ie. Hidden Markov Model)
  - Avoid Complete Speech Model
  - Pattern Classification based on Training Data

Background: Issues with Statistics

- Issues associated with “traditional” statistical ASRs:
  - Classifier trained with large variance data will not be optimized for any particular sub-problem
  - Not scalable and not easily flexible for new problems
  - Knowledge Representation is not transparent
    - Fuzzy behavior, poor re-use of knowledge: No learning
  - Optimization requires hand-crafting or “fudging” probabilities
    - Hard Coding for specific applications, conditions, and environments.

Background: Traditional ASR

- Traditional ASR Consists of:
  - Weak Model
  - Require Training Data
  - Feature Selection/Extraction
  - Pattern Classification (statistical)
  - Neighbor independent Spectral Analysis on short term time slice

  } Statistical Feature Classification based on Frequencies

- A Better Way:
  - Better Understanding and utilization of speech specific knowdedge
  - Understanding Human Speech Perception

  } Human Speech Perception Model

Background: Human Element

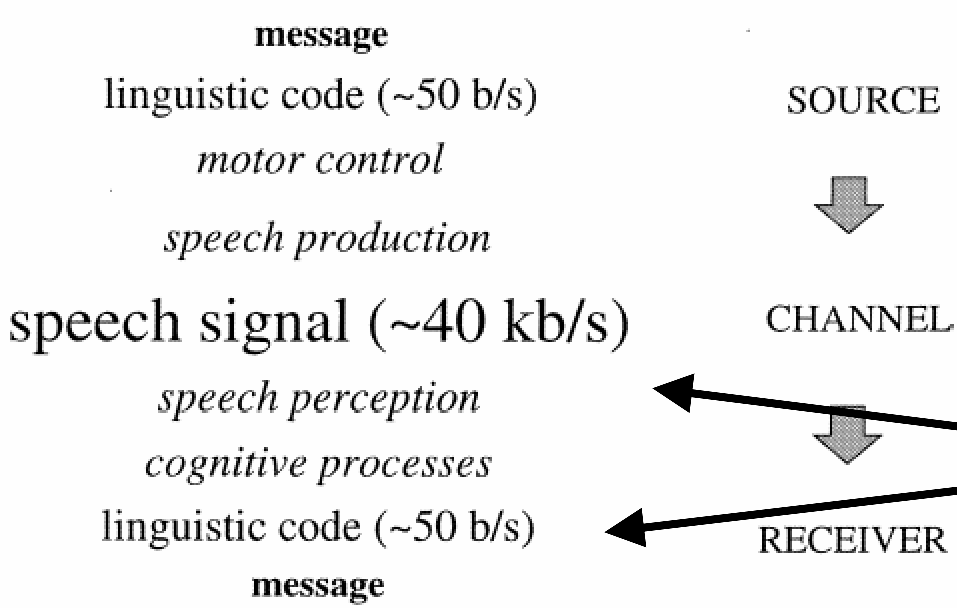


Fig. 1. The information rate in human speech communication process is highest on a speech signal level. An important role of speech perception is to reduce this rate by alleviating some of nonlinguistic variability.

Human Process optimized by forces of nature

Imperfect, Noisy

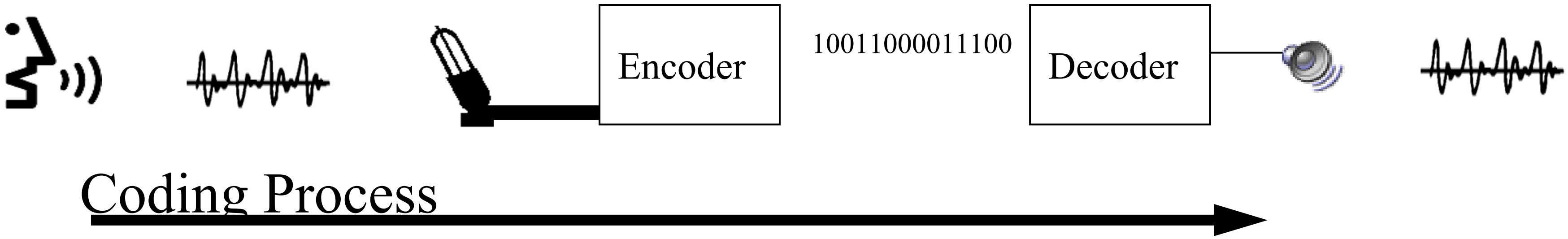
3 orders of magnitude!

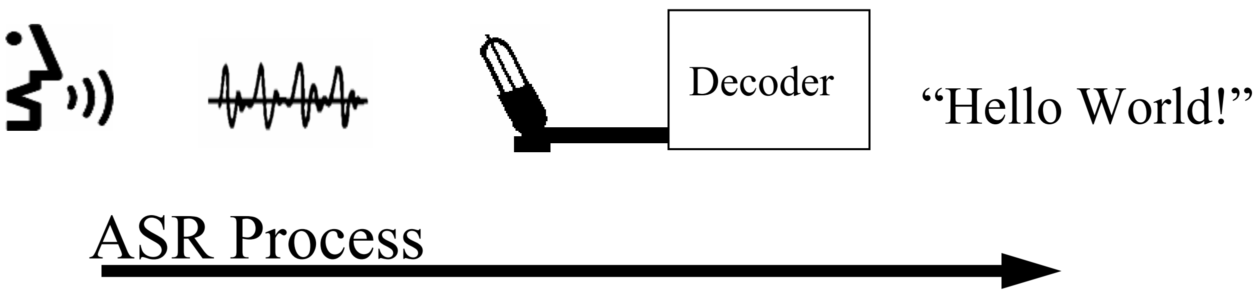
Analysis: The Signal

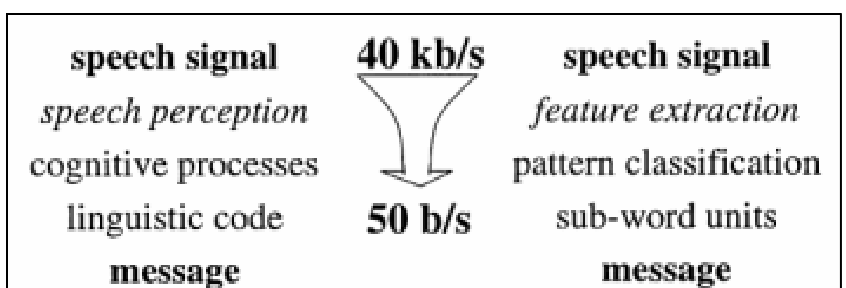
- Speech Signal Variables:
  - Vocal Tract
  - Fundamental Frequency (F0)
  - Acquired Habits (rate, accent)
  - Environment/COM Channel (noise, distortions)

Ideal ASR ignores such variables

Analysis: Coding vs. Recognition







Analysis: Approach

- Reasons for delayed progress
  - Statistical Classification is well established and understood
  - Fear of change
  - Statistical Classifiers perform well with controlled environment
  - Lack of understanding of the Speech model:
    - Dimensionality Reduction

- Approach:
  - Filter out what humans can not hear
  - Filter out noise or unneeded frequencies that do not carry msg

Spectral: Overview

- Signal Processing Algos that emulate human hearing
  - Non-Linear (Bark, Mel) Freq. Scales
  - Spectral Amplitude Compression
  - Decreasing Sensitivity of hearing at lower freq. (equal-loudness)
  - Large Spectral Integration by:
    - PCA
    - Ceptstral Trucation
    - Low order autoregressive modeling

Spectral: _Linear Prediction_

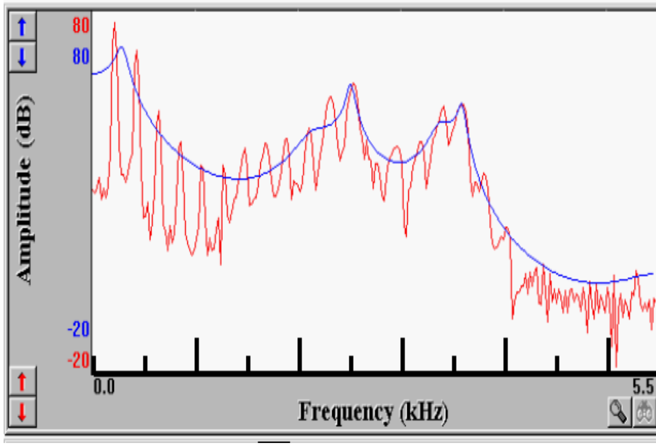
"mathematical operation where future values of a discrete-time signal are estimated as a linear function of previous samples."

$$\hat{x}(n) = -\sum_{i=1}^{p} a_i x(n-i)$$ Prediction

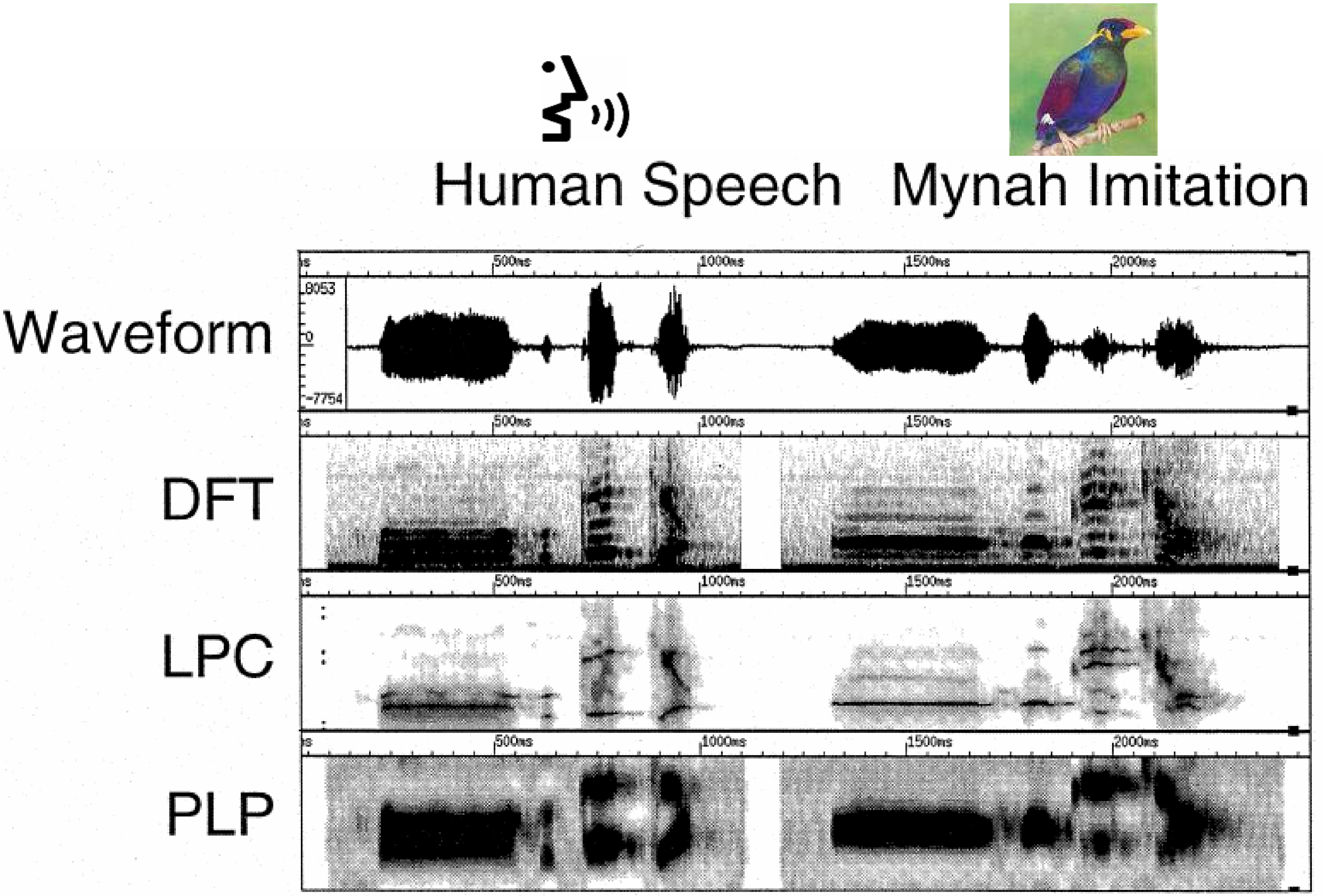
$$e(n) = x(n) - \hat{x}(n)$$ Error

- Linear Predictive Coding (LPC)
  - Compressed Representation of the *Spectral* Envelope

- Perceptual Linear Prediction (PLP)
  - *Human characteristics* applied to engr. approximations



Spectral: PLP

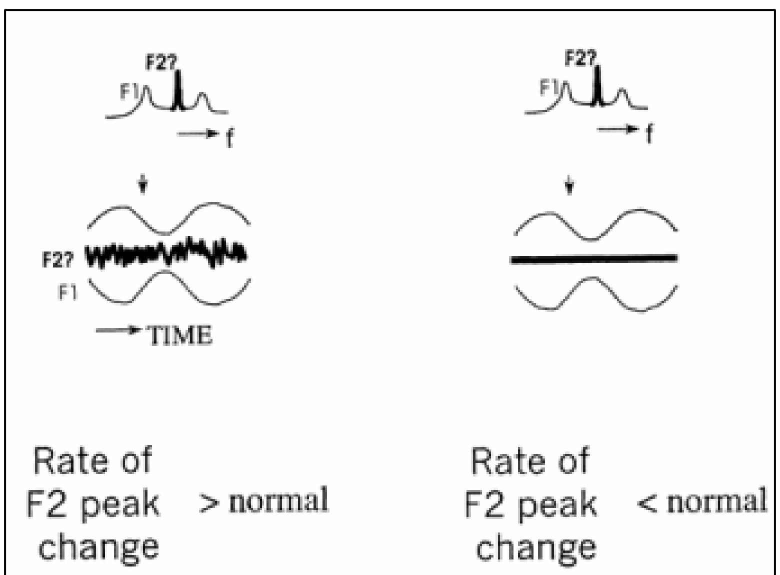


Spectral: level of detail

- Message is in gross spectral features
  - With low-detail spectrum:
    - ASR performs better on cross-speaker data
    - Speaker dependent information is minimized
- Revisit notion of formant significance
  - Humans do not resolve higher formants
  - Focus on positions and shapes of whole formant clusters to extract linguistic message

Temporal: Overview

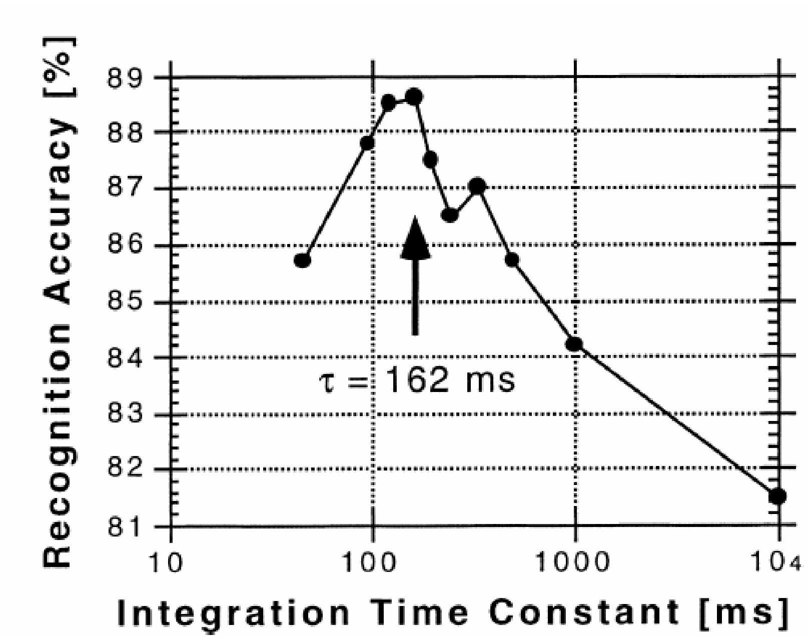
- Traditional ASR (ie. HMM) assume signal as short (10-20ms) steady-state segments.
  - Each segment is represented by a vector classified as phoneme
  - Issues with short segmenting: - CONTEXT
    - Coarticulation, forward masking, syllables, noise



"zoom"

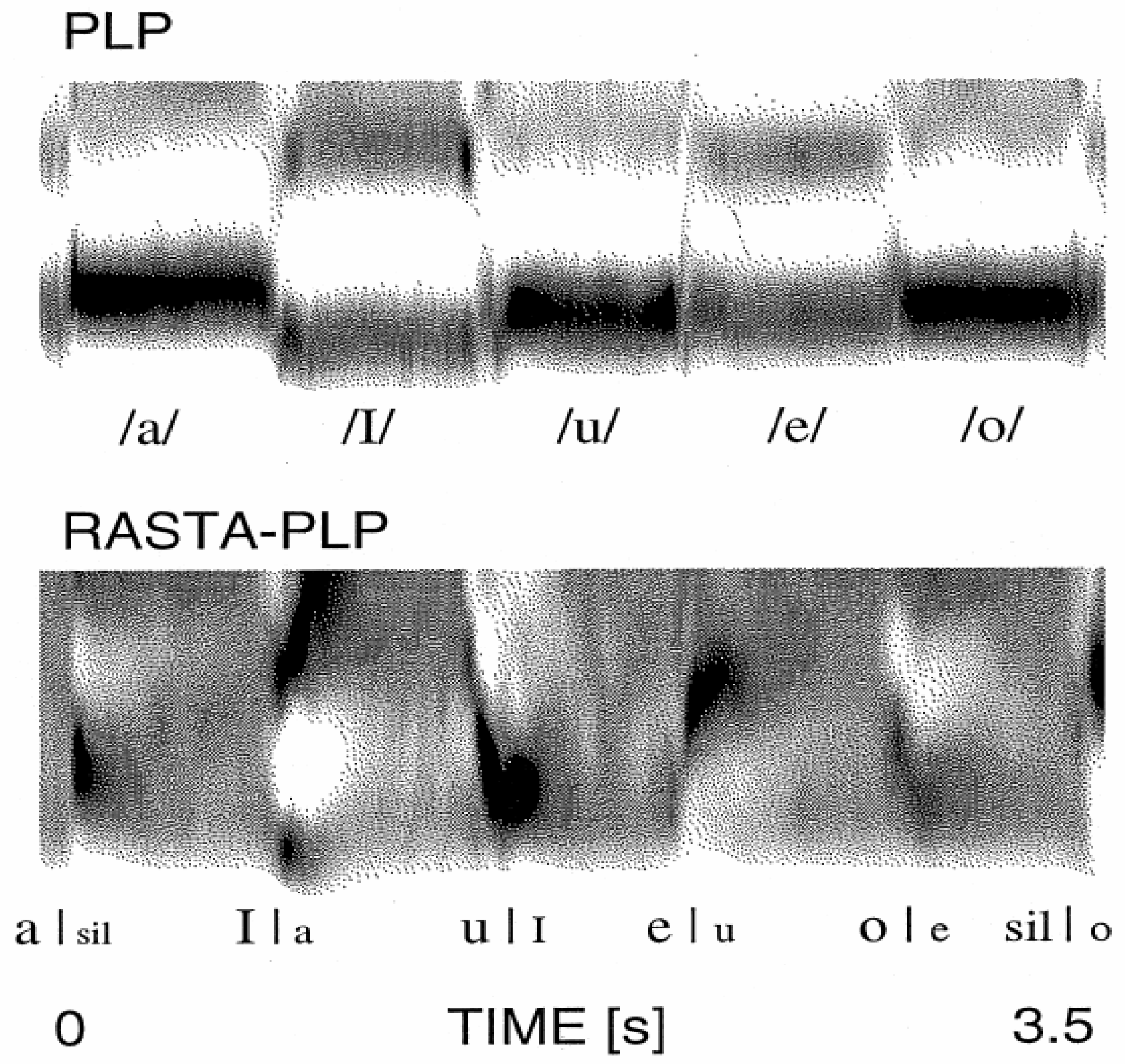
Temporal: RASTA (1)

- RelAtive SpecTrAl (RASTA)
  - Removes fixed (slow varying) nonlinguistic components of speech features
  - Assumes “fixed” noise through time in speech
  - RASTA band-pass filtering is done on the log spectrum



RASTA Optimization

Temporal: RASTA (2)



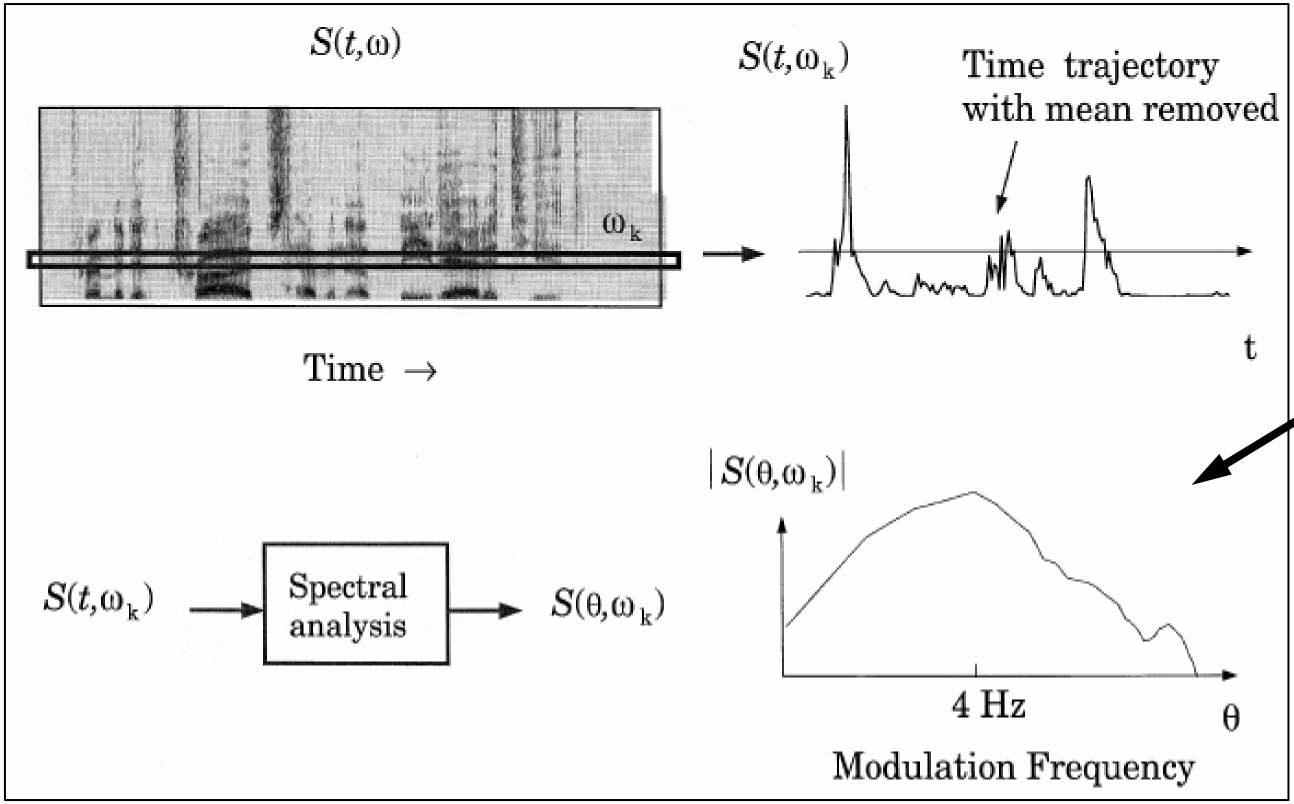
Since RASTA Removes slow-varying features (noise),

<u>RASTA emphasizes changes in the signal</u>

Temporal: Modulation

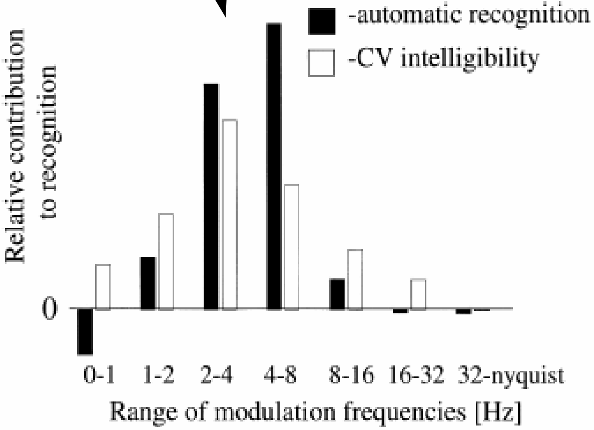
Primary carrier of linguistic info are <u>changes</u> in the vocal tract

Changes are reflected in the spectral envelope



2-8 Hz represent the phonetic rate in speech

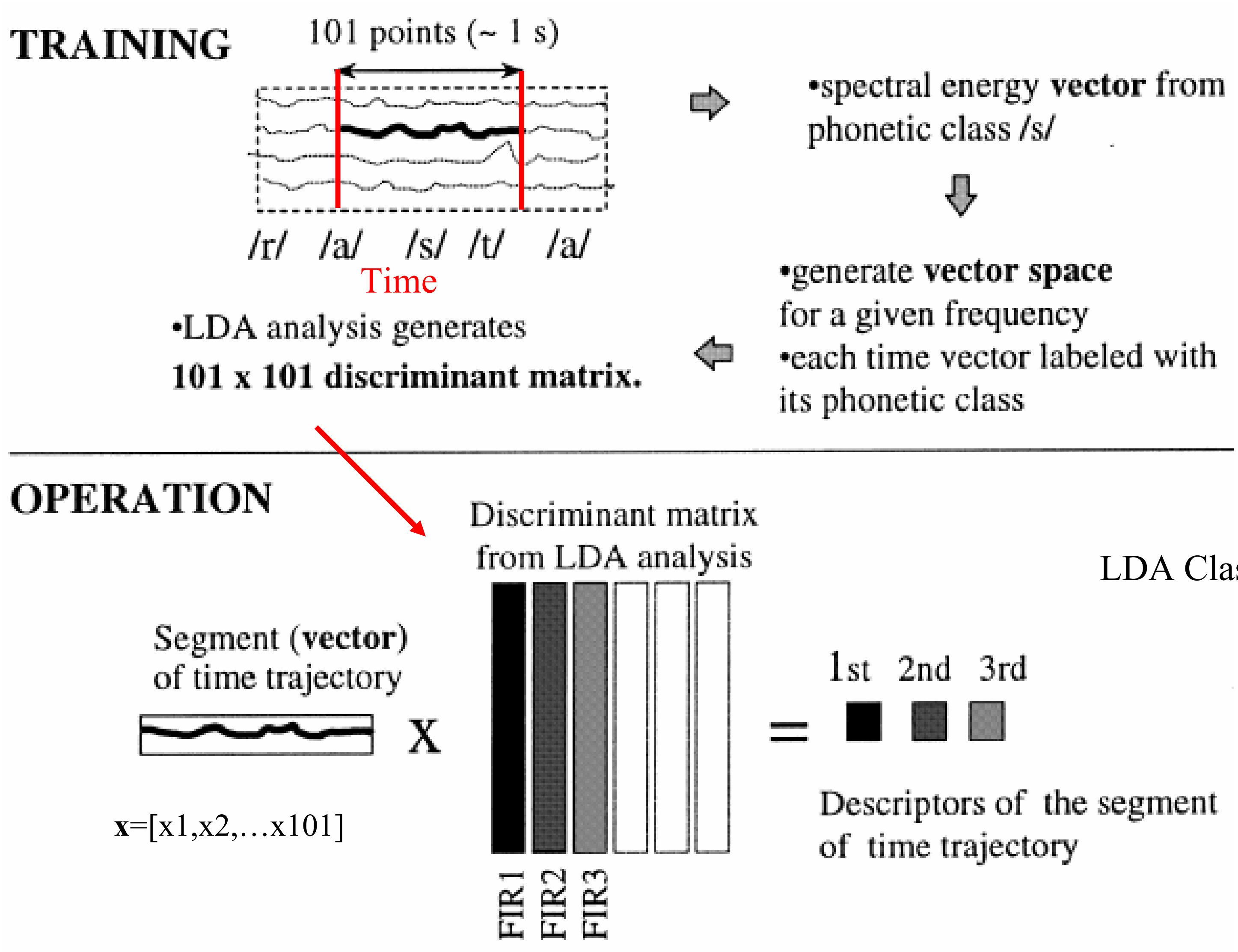
(~150-250 ms)



Temporal: Data-Driven RASTA (LDA)

Input vector: segments of time trajectory of a single log critical-band energy over 1s time-span

Fs = 100 Hz



LDA Classifier

Finite Impulse Response

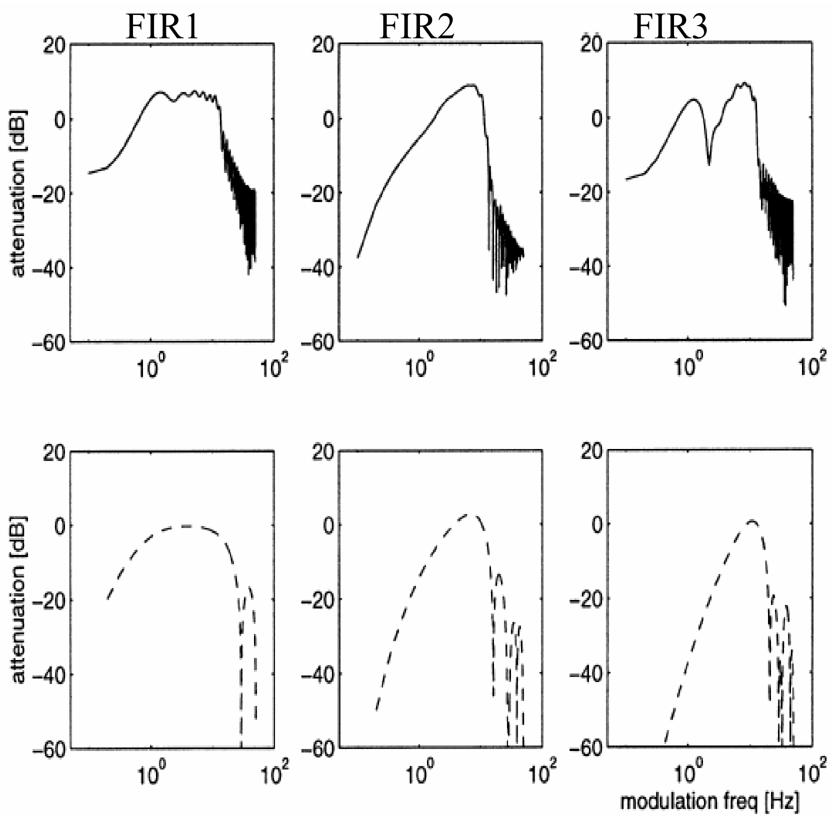
Temporal: Data-Driven RASTA (Test)

Test Data is 30 min of hand-labeled phone conversations using critical band centered at 5 Bark (450 Hz).

Frequency Characteristics

First 3 LDA discriminant vectors

RASTA Filtered

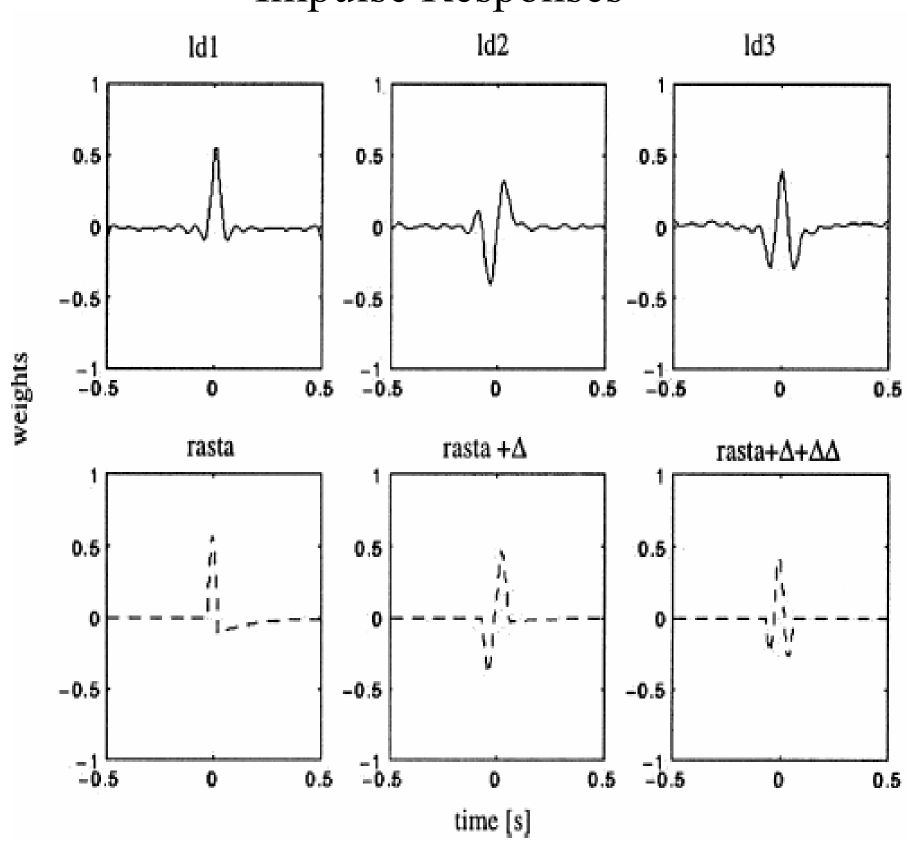


Original

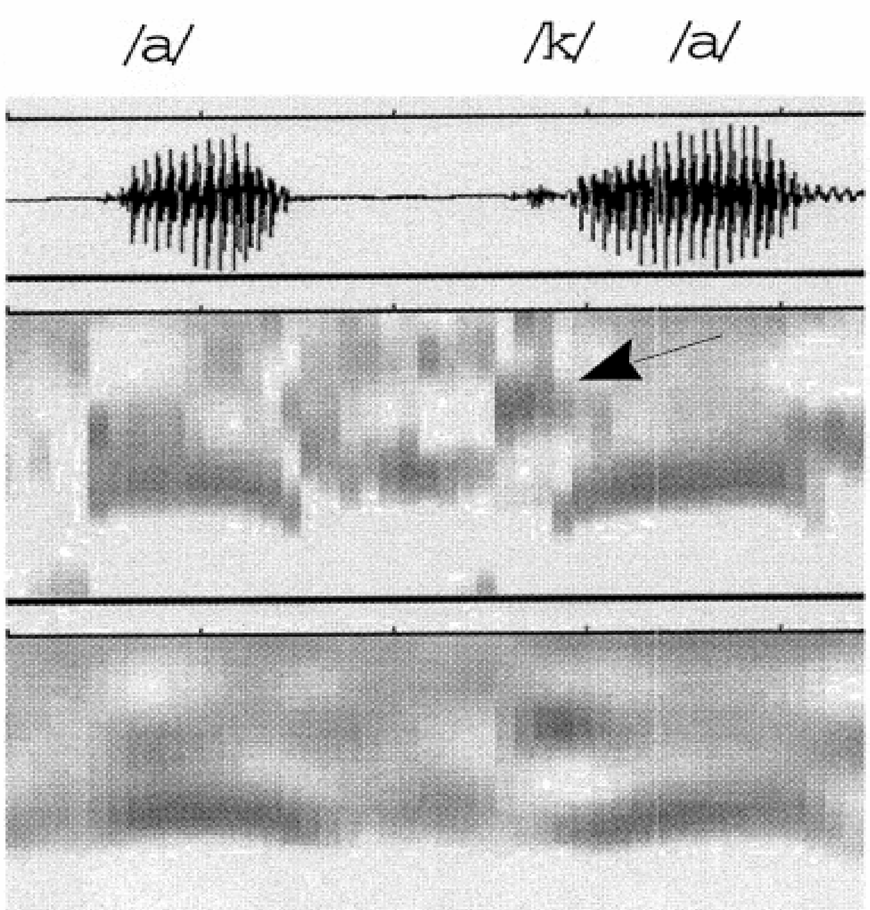
2^{nd} orthogonal polynomial (slope) over 90ms

3^{rd} orthogonal polynomial over 90ms (curvature)

Impulse Responses



Temporal: RASTA sluggishness



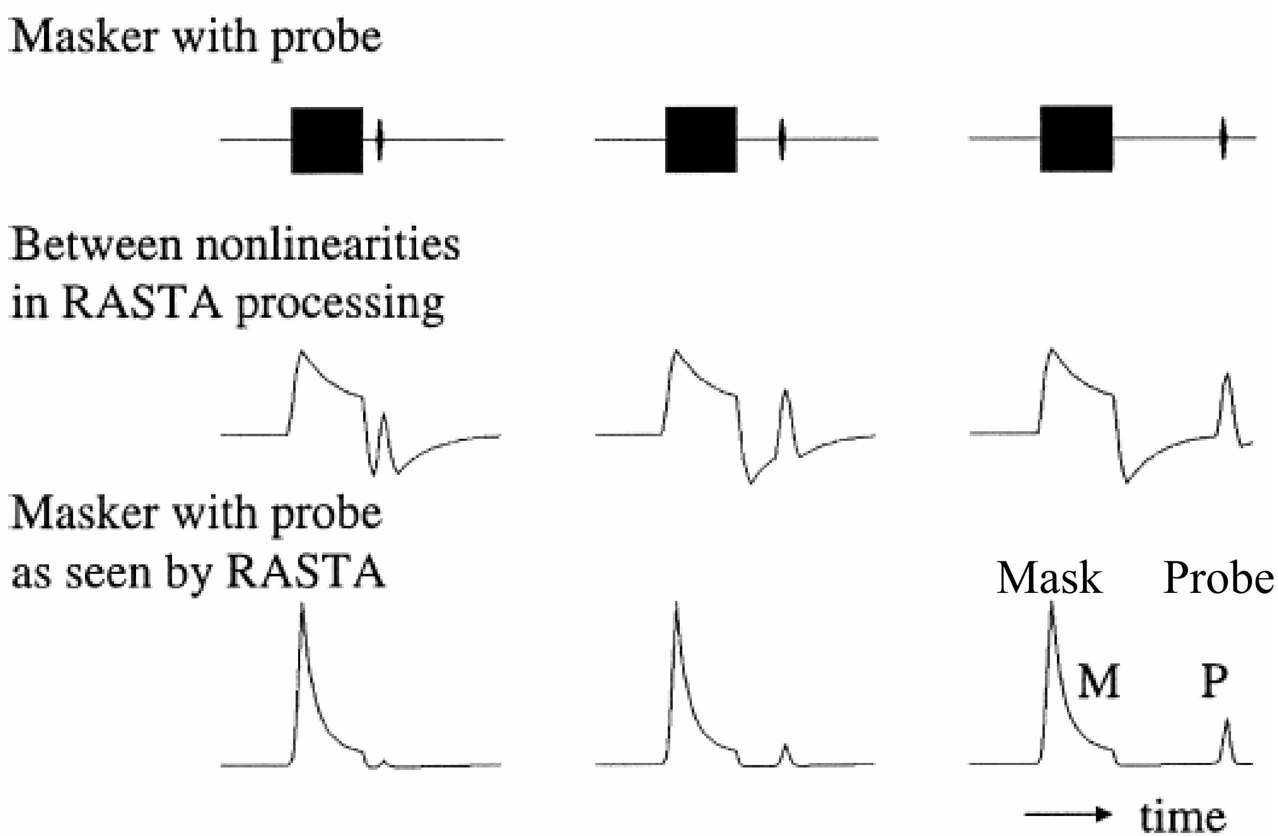
PLP

Traditional short 10ms segments

RASTA-PLP

Sluggish 200 ms resolution

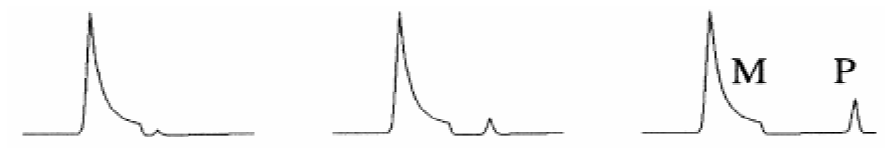
Temporal: Masking (1)



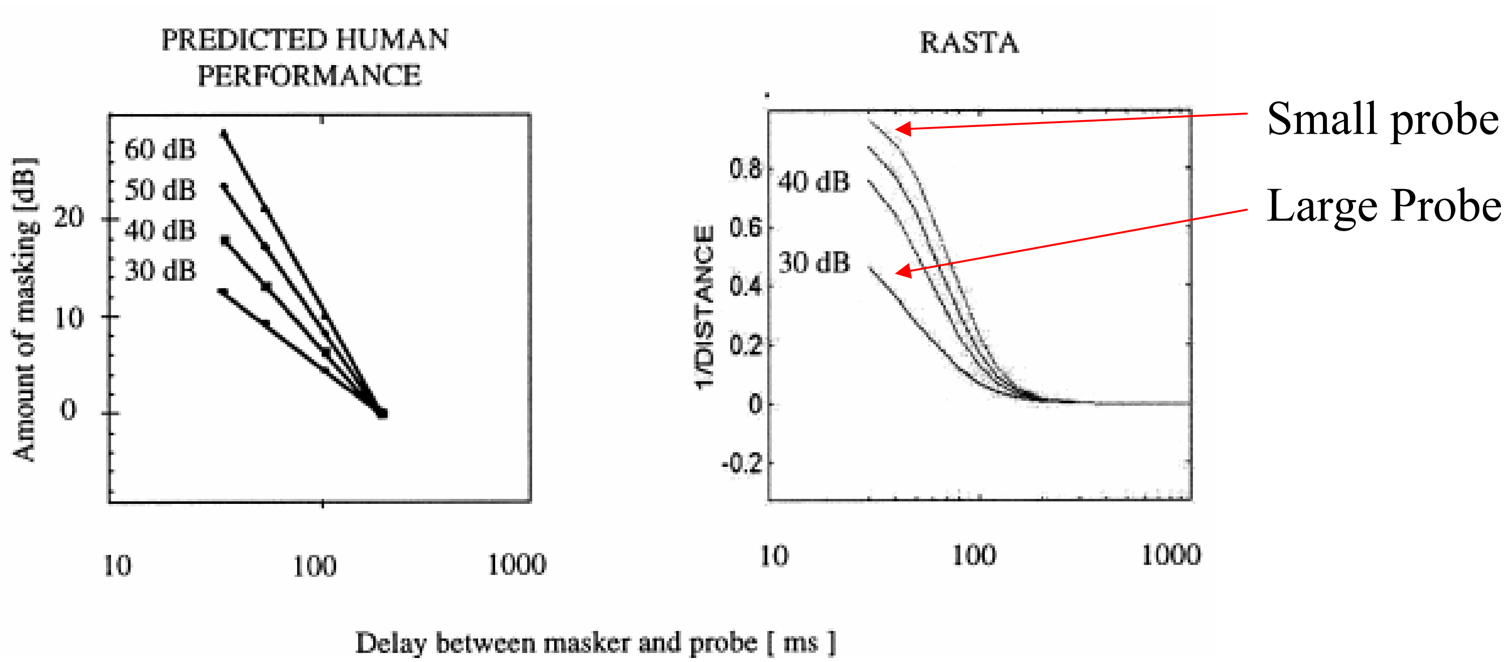
-Rasta Emulates Human Perception.

-PLP would not catch the masking effect.

Temporal: Masking (2)

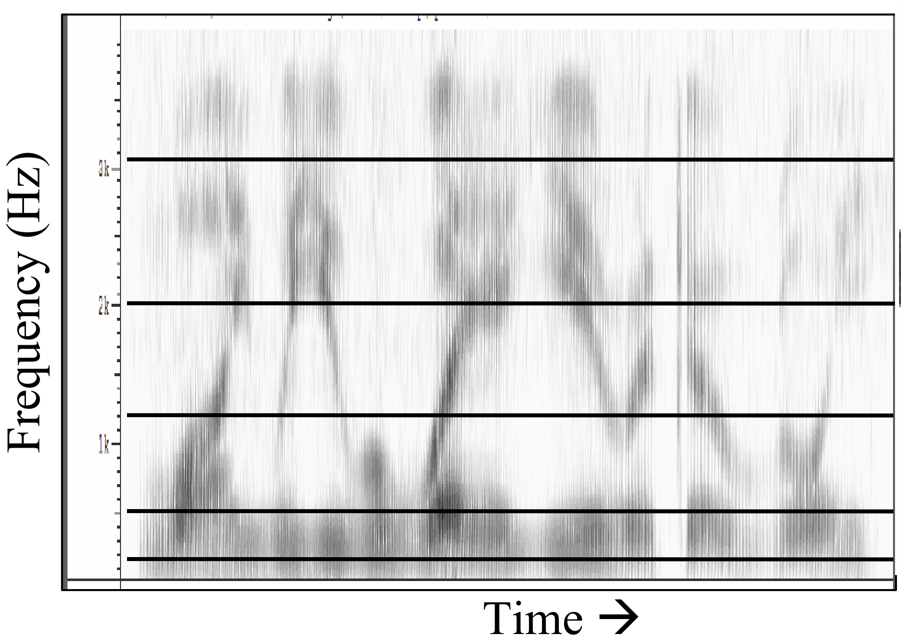


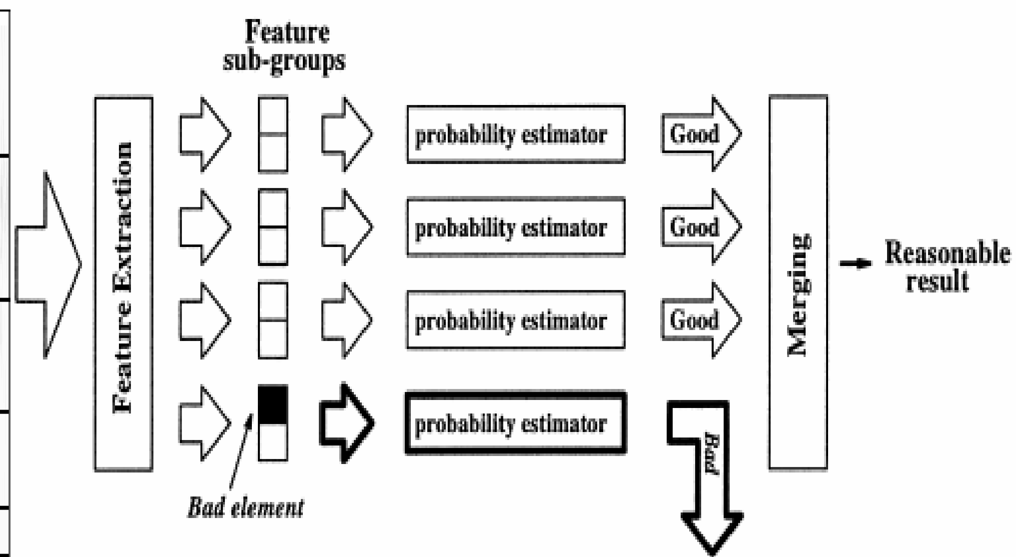
small probe	Large Probe
small distance	Large distance (vs no probe)
More Masking	less Masking



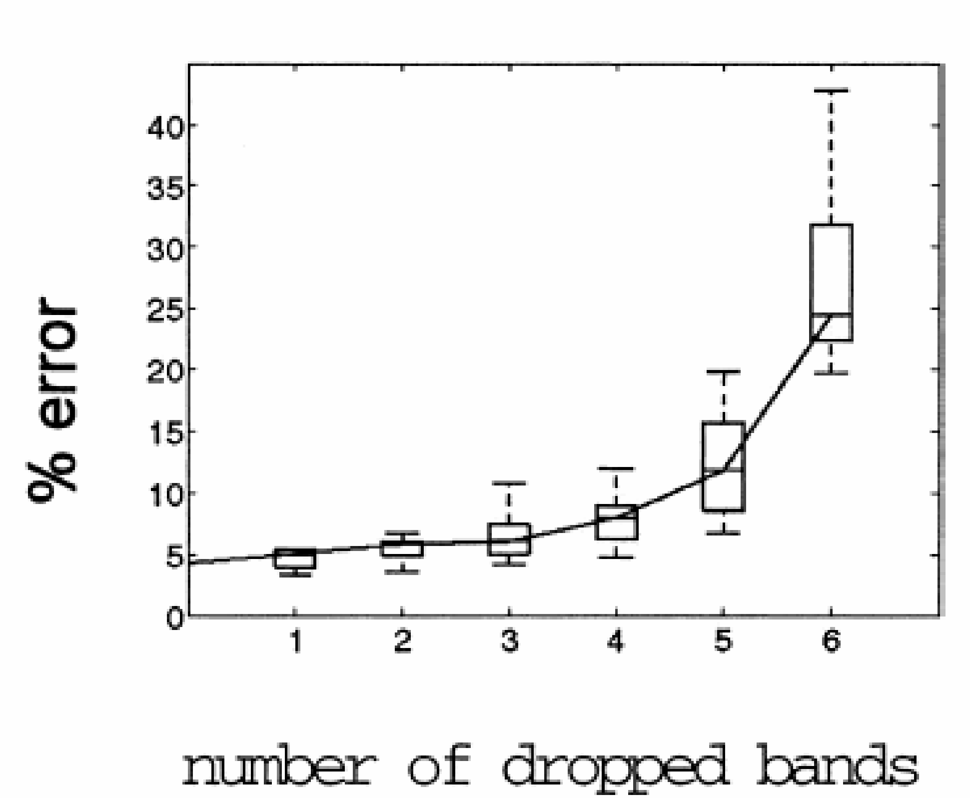
Partial Information (1)

- Speech signal is easily corrupted or distorted by noise
- Noise has minimal effect on human perception
- Humans split the signal in sub-bands (redundant information in each sub-band)
  - Then decode individual sub-bands, drop bands with high noise
  - Reliable information from one sub-band is sufficient to discard others





Partial Information (2)



Slow ASR degradation when omitting information from the signal: Verifies Redundancy

Conclusions

- Perception is decoding linguistic message
- Understanding the human speech model is required
- Use and design for "real speech data"
- Discourages traditional pattern-matching approach
- Speech contains noise and excessive data that provides no useful information