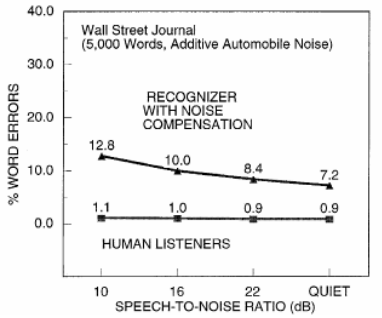
Cognitive Modeling & Processing for Speech Recognition – Ears and Beyond

B.H. Juang & Woojay Jeon
Georgia Institute of Technology

February 2, 2005

Introduction

- ASR systems still perform far worse than human listeners under noisy conditions.



- Where is the bottleneck?
  - Syntactics & semantics – the language issues
  - **Acoustics & phonetics – the auditory**/articulatory **issues**

Overview

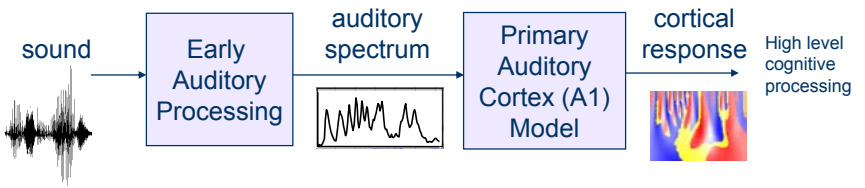
- Premises: Ear and brain need to work together.
- While approximations of the peripheral system have been considered in the past, models of the primary auditory cortex (A1) have been scarcely used in previous studies, particularly related to ASR.
- We experimentally establish the relevance of the **auditory model** (auditory spectrum + cortical response) in the existing speech recognition framework (which is mostly mathematical).
- We relate auditory modeling to current common practices in feature selection (e.g., MFCC as a crude approximation to the auditory system) – how far can we go if we can afford the sophistication.

  ➔ **Lessons, insights, new directions**

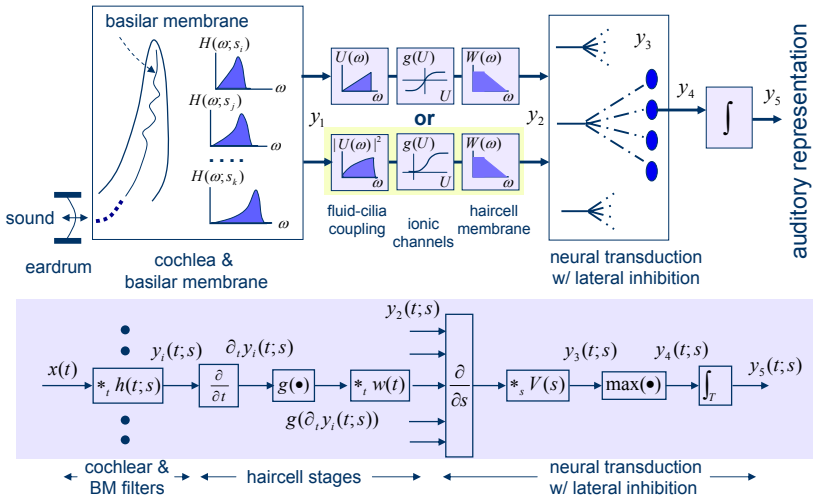
Perceptual Processing of Audio Signals

We chose the computational model of Yang & Shamma [1992] for the peripheral and that of Wang & Shamma [1995] for the "central" auditory system as a starting point in our investigation.

Overall Model

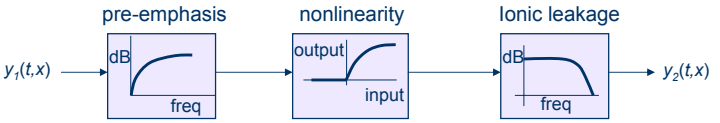


Early Auditory Processing (Yang & Shamma)



Inner Hair Cells

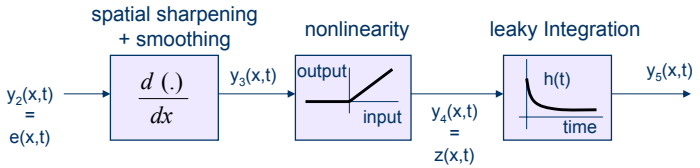
- The inner hair cells convert filter outputs to electrical activity along a tonotopically ordered nerve array, usually modeled by a 3-step process:



  - Pre-emphasis stage (HPF): coupling of fluid velocity and hair cell cilia, modeled by a temporal derivative.
  - Hair Cell Nonlinearity: limits the dynamic range.
  - Ionic Channel Leakage: gradually attenuates the signal response beyond 4-5kHz.

LIN Reduction in the Auditory Spectrum

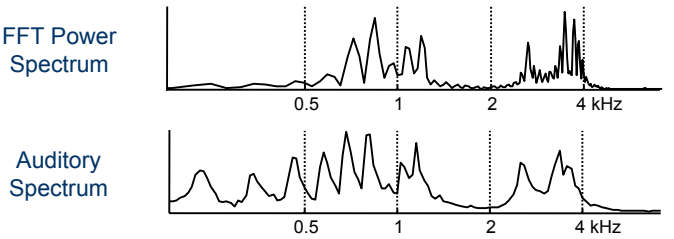
- Hence, LIN reduction at the cochlear nucleus can be implemented by the following three stages:



  - Inhibition influence among proximate neurons: derivative across the channels.
  - Neuron Threshold Nonlinearity: half-wave rectifier.
  - Temporal Smoothness: slow dynamic of neurons; modeled by a leaky integration across time.

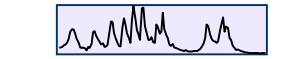
Auditory Spectrum

- The output is called the “auditory spectrum” due to similarity to the STFT spectrum but with non-linearly transformed frequency and gain – this internal representation is used as the input to the **primary auditory cortex (A1)** model.
- Data is highly compressed, but retains most auditory information, e.g. pitch, timbre, voice quality.
- Analytically shown to be self-normalizing and noise-robust.



Central Auditory Processing – Shamma's A1

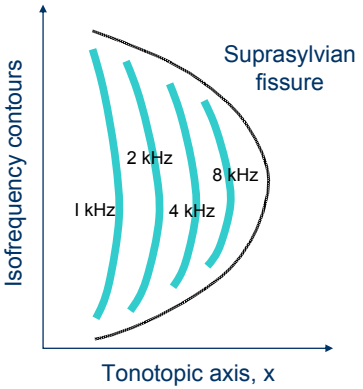
Auditory spectrum



Central Auditory Processing

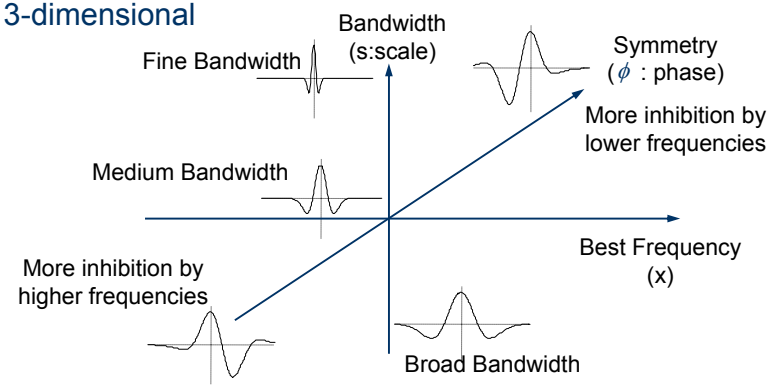
Response at A1
(3-D Cortical Response)

Response at primary auditory cortex A1



Organization of Response Areas in the A1

3-dimensional

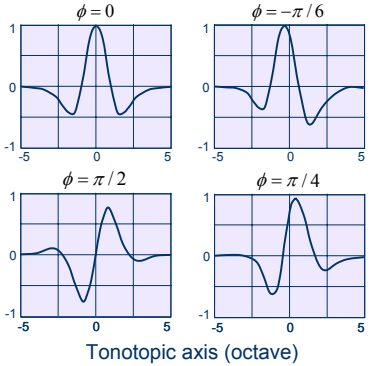


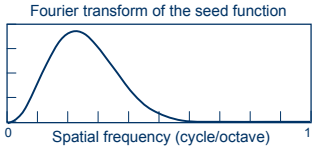
- Each response area is defined on the frequency axis y and parameterized by three variables: x (Best Frequency), s (Bandwidth), ϕ (Symmetry)

$$\rightarrow w(y; x, s, \phi)$$

Asymmetry Response Model

- Response function is modeled using sinusoidal interpolation between a seed function $h_s(x)$ and its asymmetric counterpart (Hilbert Transform) $\hat{h}_s(x)$.





Bandwidth Scaling: $H_s(k) = H_m\left(\frac{k}{\alpha^s}\right)$

Response unit:

$$w_s(y; x, s, \phi)$$
$$= h_s(y-x)\cos\phi - \hat{h}_s(y-x)\sin\phi$$

Calculation of Cortical Response

- The neural response function with best frequency x, symmetry index ϕ and scaling index s is modeled as:

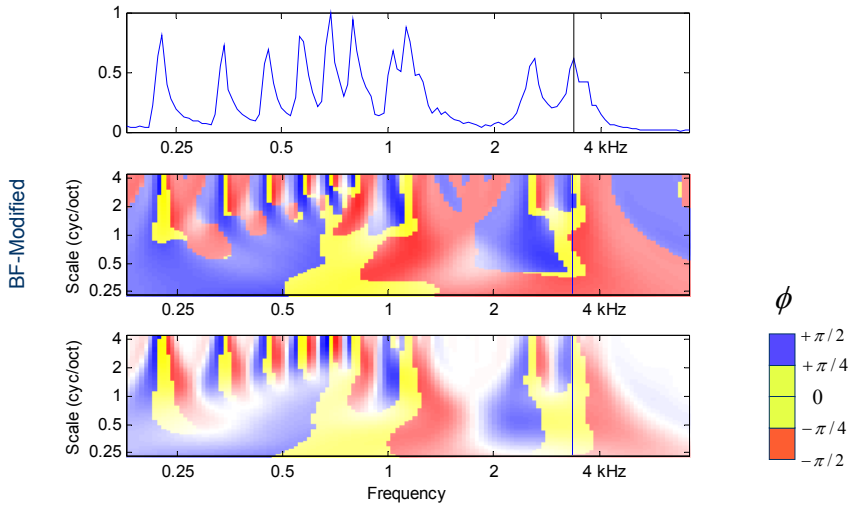
Response unit:

$$w_s(y; x, s, \phi) = h_s(y-x)\cos\phi - \hat{h}_s(y-x)\sin\phi$$

- The response to a given input spectrum $p(y)$ is then calculated by taking the inner product between the input and the different response functions:

$$r_s(x, s, \phi) = \langle p(y), w_s(y; x, s, \phi) \rangle_y = \int_R p(y) w_s(y; x, s, \phi) dy$$

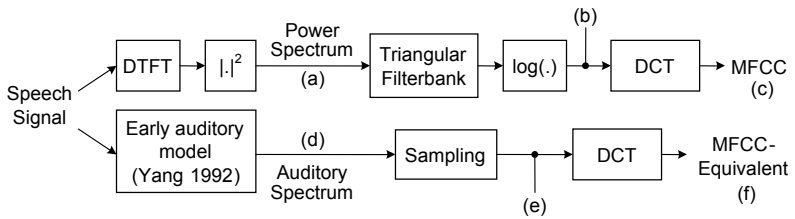
Example of Cortical Response



Relating Cortical Response with Others

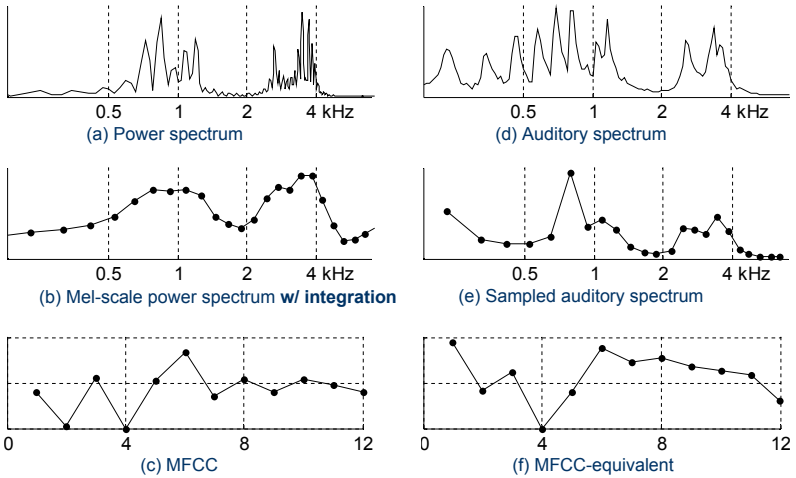
- If this (refined) cortical representation is justifiable from physiological perspectives, how does it compare to known practices, say the MFCC? Starting with the auditory spectrum?
- Can any improvements on the MFCC be found? (After all, the "optimality" of the MFCC as an approximation to auditory response was never seriously discussed.)
- Study of other non-cognitive perceptual effects.

The Auditory Spectrum and the MFCC



- The MFCC is mostly based on a crude approximation of the peripheral auditory system, most notably the cochlear filtering action where spectral energy is integrated.
- A crude counterpart to the MFCC could be extracted from the auditory spectrum using the method shown above – mainly aligning BF's with the CF's of the MFCC filterbank.

The Auditory Spectrum and the MFCC

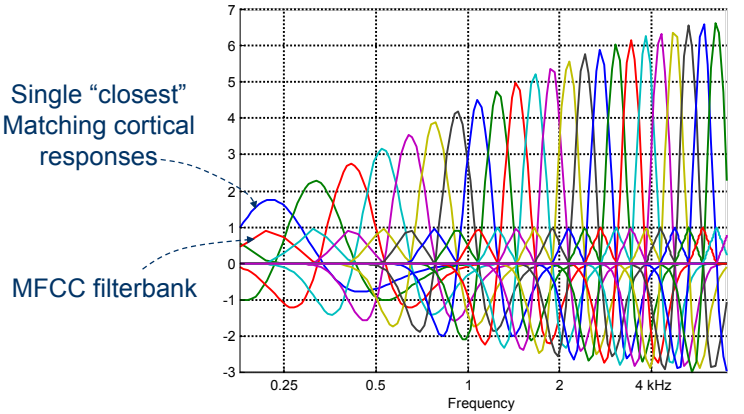


Speech Feature Extraction from the A1

- $\mathbf{r}_s$ contains too much data (128x21x11 = 29,568 points), more than we can use or understand at the moment
- In comparing to MFCC (12-dimensional) for use in speech recognition, apply PCA and LDA to the cortical response to derive 12-dimensional feature vectors
  - Initial dimensions too many; retain 2,000~3,000 points where variance is smaller;
  - Apply PCA to reduce the dimension further down to 40;
  - Apply LDA (***preliminary***, limited by size of dataset).

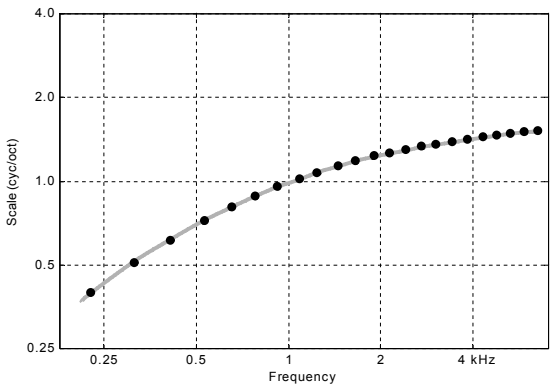
A "Cortical Cepstrum" (1)

- We find cortical response areas that correspond to the MFCC filterbanks in terms of center frequencies and bandwidths.



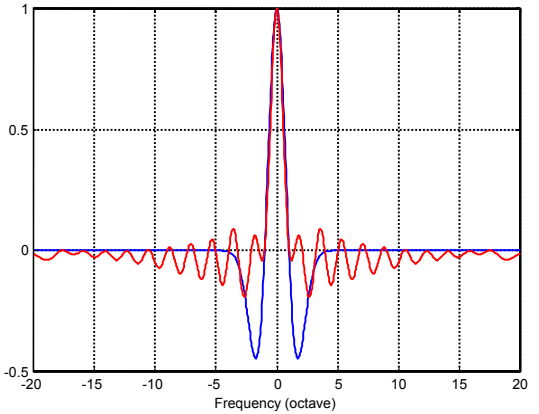
A "Cortical Cepstrum" (1)

- Each response area can be indicated by a dot on the zero-phase cortical plane, conceptually demonstrating that the filterbanks constitute a subset of the cortical response.



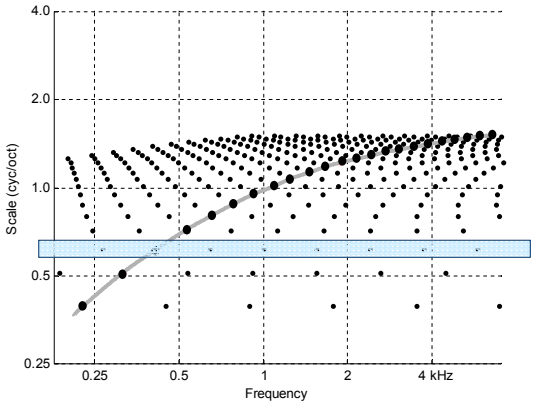
A "Cortical Cepstrum" (2)

- To avoid data loss at the inhibitory regions, we try taking a linear combination of a set of response areas to better simulate the integration.



A "Cortical Cepstrum" (2)

- Each triangular filter response is a linear combination of a set of cortical response points located on a horizontal line surrounding the center response area (the big dark dot).



Principle Component Analysis

- From a set of speech training data, we take numerous samples of the vector $\mathbf{r}_s$ and compute the scatter matrix:

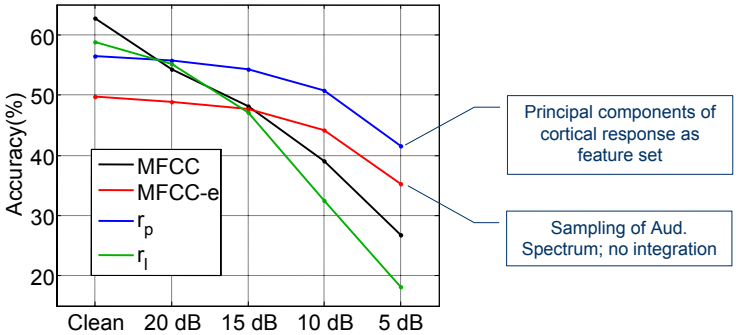
$$\mathbf{S} = \sum_{\mathbf{r}_s \in C} (\mathbf{r}_s - \mathbf{m})(\mathbf{r}_s - \mathbf{m})^T$$

where m is the sample mean of the entire training data.

- Taking the eigenvectors of $\mathbf{S}$ with the p largest eigenvalues and arranging them as the columns of the matrix $\mathbf{E}$, we can compute the p-most principle components of $\mathbf{r}_s$ by:

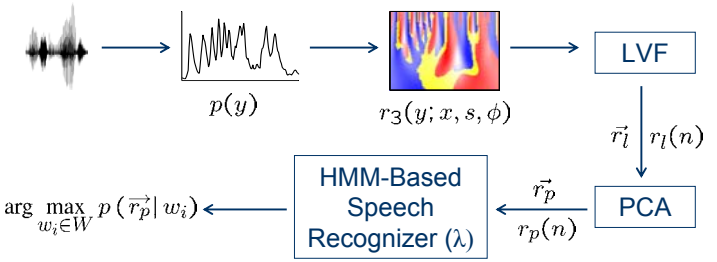
$$\mathbf{r}_p = \mathbf{E}^T(\mathbf{r}_s - \mathbf{m})$$

Phoneme Recognition Results



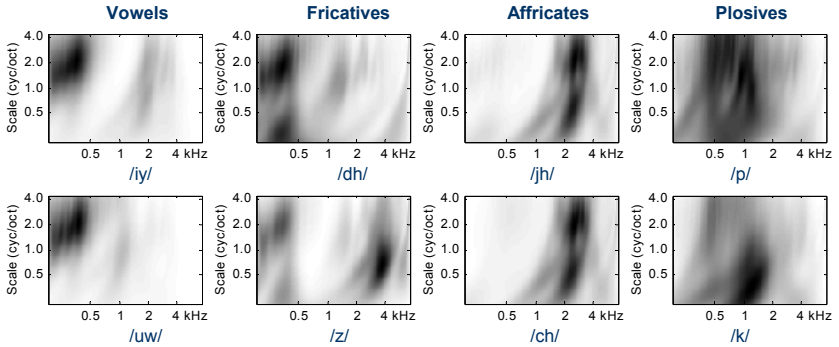
- 38 phonemes segmented from TIMIT were used for the recognition task.
- Results show that features derived from auditory model yield results comparable to those of MFCC.
- Noise robustness of auditory spectrum contribute to better performance of MFCC-e and r_p.

Traditional Recognizer



- In the previous recognition task, we applied a single **low variance filter** to all phoneme samples, ignoring the category-specific low variance regions.
- Category-independent PCA, after transformation, may obscure the meaning of the original responses.

Low Variance Regions



- Regions in the cortical response with **low variance** (light areas above) are likely to contain the identifying features of each phoneme.
- Results imply that phonemes sharing common characteristics may share common low variance regions.
- Place coding?

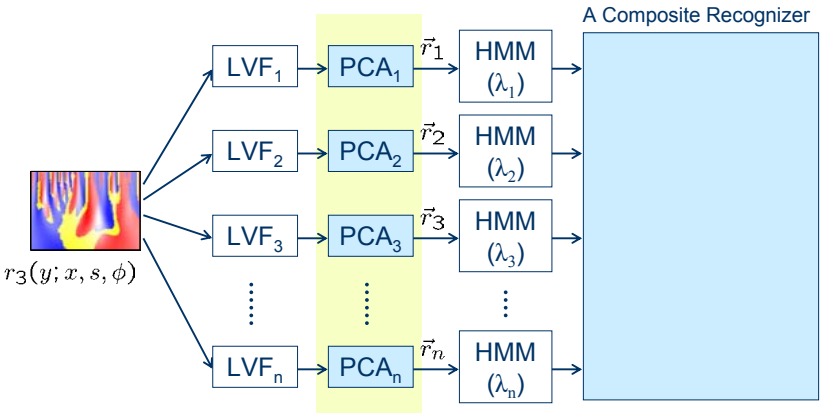
Cognition-Based Feature Selection

- Auditory features are selected based on:
  - Perceptual place-coding phenomena observed in the physiological response
    → How should the phoneme classes be clustered?
  - A hierarchical cognitive process we conjecture to take place in the auditory system
    → How should the category-to-feature dependencies be modeled?

Phoneme Categorization

1. Front vowels: iy, ih, ey, eh, ae
2. Mid vowels: er, ah
3. Back vowels: uw, uh, ow, aa
4. Voiced fricatives: v, dh, z
5. Unvoiced fricatives: f, th, s, sh
6. Whisper: hh,
7. Affricates: jh, ch
8. Nasals: m, n, ng
9. Diphthongs: ay, aw, oy
10. Liquids: r, l, dx
11. Glides: w, y
12. Voiced consonants: b, d, g
13. Unvoiced consonants: p, t, k

A Simple Recognizer w/ C-D Observations



- Employ n low-variance filters obtained from n phoneme categories to produce n features and n intermediate recognizers.

A Composite Recognizer

- Many combination rules exist for multiple recognizers.
- One simple method is to use a MAP decision rule with the assumption that observations in subspaces are conditionally independent:

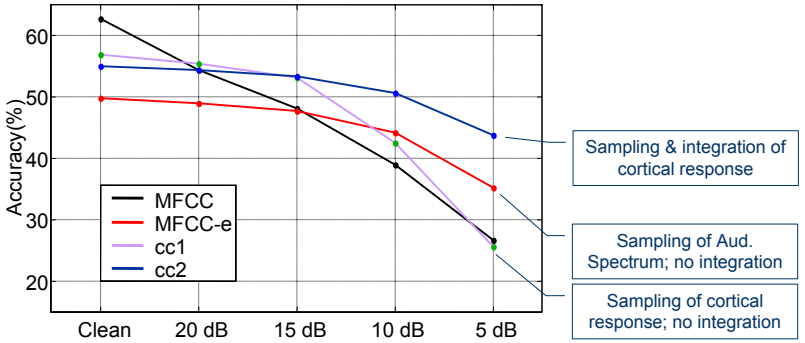
$$p(\vec{x}_1, \ldots, \vec{x}_n | w_i, \Lambda) = \prod_{j=1}^{n} p(\vec{x}_j | w_i, \Lambda)$$

where $\Lambda \triangleq \{\lambda_1, \ldots, \lambda_n\}$

Then, $\arg\max_{w_i \in W} p(w_i | \vec{x}_1, \cdots, \vec{x}_n, \lambda_1, \cdots, \lambda_n)$

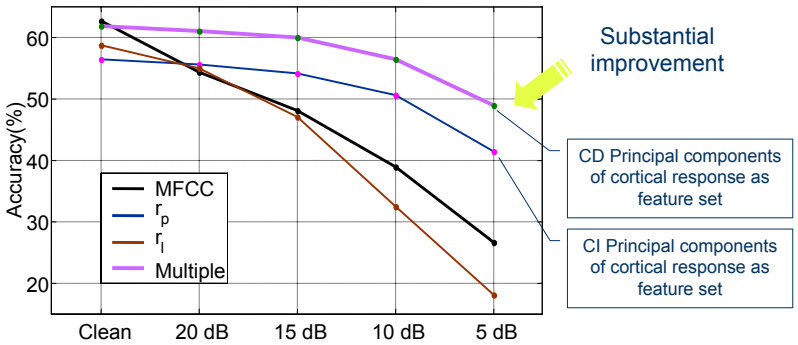
$= \arg\max_{w_i \in W} \sum_{j=1}^{n} p(\vec{x}_j | w_i, \lambda_j)$ assuming uniform prior

Phoneme Recognition Results (1)



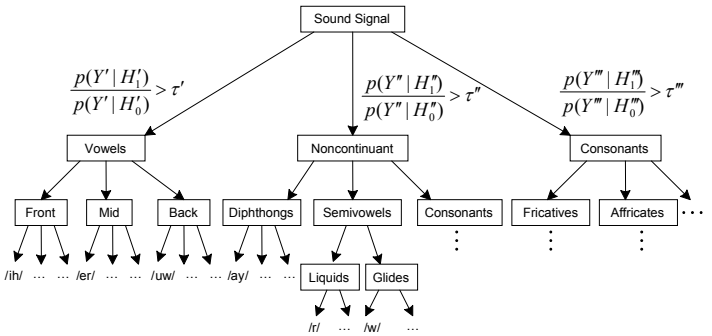
- MFCC-e: MFCC-equivalent feature from **auditory spectrum**.
- cc1: Cortical Cepstrum Type 1 (one-to-one)
- cc2: Cortical Cepstrum Type 2 (integrated)

Phoneme Recognition Results (2)



- r_p: PCA-derived features from cortical response
- r_l : LDA-derived features from cortical response - preliminary
- "Multiple": Results using 13 different LVF's and 13 recognizers based on phoneme categorization.

Hierarchical Detection



- Phoneme hierarchy
- Clustering techniques may be employed to obtain "better" hierarchical structures.
- Similar structure may be used for non-speech audio signals.

Future Work

- Further develop the relationship between the MFCC and the cortical response to gain more insight on how to derive refined auditory features.
- A more rigorous development of *"identifying region"* and place-coding in the cortical response.
- A detection-based, hierarchical framework for perceptual and cognitive analysis