A Study of Auditory Modeling and Processing for Speech Signals

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

Nov. 4, 2004

Objective

- Understand auditory models
- Experimentally establish the relevance of auditory model (auditory spectrum + cortical response) in the existing speech recognition framework (which is mostly mathematical)
- Relate auditory modeling to current common practices in “feature selection” (e.g., MFCC) – how far can we go if we can afford the sophistication
- Study the cortical response in connection with known physiological studies to gain better understanding of auditory physiology, which inspires us to develop hierarchical, detection-based methods in the future

Perceptual Processing of Audio Signals

Early Auditory Processing

Overall Model

Bio-mechanical transduction - The Ear Model
mostly physical and mechanical

Bio-chemical Transduction
mostly chemical, some physical

Early Auditory Processing

sound → Outer & Middle Ear → Cochlear (Inner Ear) → Inner Hair Cell → Neural Networks (w/ Lateral Inhibition) → auditory spectrum → Primary Auditory Cortex Model → High level cognitive processing → auditory representation
**Inner Hair Cell**

- Inner Hair Cells convert filter outputs to electrical activity along a tonotopically ordered nerve array, usually modeled by 3 processes:
  - Pre-emphasis stage (HPF): coupling of fluid velocity and hair cell cilia, modeled by the temporal derivative.
  - Hair Cell Nonlinearity: louder signal induces higher firing rate.
  - Ionic Channel Leakage: gradually attenuates the signal response beyond 4-5kHz.

**Neural Networks with Lateral Inhibition**

- Transformation of the fine temporal responses to spatial variation at cochlear nucleus, consisting of 3 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 “auditory spectrum” due to similarity to STFT spectrum but with non-linearly transformed frequency and gain – the input to auditory cortex.
- Data is highly compressed, but retain most auditory information, e.g. pitch, timbre, voice quality.

**Central Auditory Processing – Shamma’s A1**

- Auditory spectrum
- Response at primary auditory cortex A1
- Isofrequency contour
- Tonotopic axis, x
- Suprasylvian fissure
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), $\phi$ (Symmetry)

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

Asymmetry Response Model

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

Response unit:

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

Calculation of Cortical Response

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

Response unit:

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

- 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:

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

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 MFCC? Starting with the auditory spectrum?
- Can improvement on MFCC be found? (After all, optimality of 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 BFs with the CFs of MFCC filterbank.

The Auditory Spectrum and the MFCC

Speech Feature Extraction from the A1

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

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. 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 $r$, and compute the scatter matrix:
  
  \[ S = \sum_{r_i \in C} (r_i - \mu)(r_i - \mu)^T \]

  where $\mu$ is the sample mean of the entire training data.

- Taking the eigenvectors of $S$ with the $p$ largest eigenvalues and arranging them as the columns of the matrix $E$, we can compute the $p$-most principle components of $r$ by:
  
  \[ r_p = E^T (r - \mu) \]

**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?
**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:

\[
p(x_1, \ldots, x_n | w, \Lambda) = \prod_{j=1}^{n} p(x_j | w, \lambda_j)
\]

where \( \Lambda \triangleq \{\lambda_1, \ldots, \lambda_n\} \)

Then,
\[
\arg\max_{w \in W} p(x_1, \ldots, x_n | w, \Lambda) = \arg\max_{w \in W} \sum_{j=1}^{n} p(x_j | w, \lambda_j)
\]

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 in [Deller, 2000].
Hierarchical Detection

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

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