Auditory Models for Speech Analysis

Jun Hou, Lawrence Rabiner, Sorin Dusan
Rutgers University
Why Auditory Models

• Match human speech perception
  – Non-linear frequency scale – mel, Bark scale
  – Spectral amplitude (dynamic range) compression – loudness (log compression)
  – Equal loudness curve – decreased sensitivity at lower frequencies
  – Long spectral integration – “temporal” features
  – Auditory masking – loud tones (or noise) mask adjacent signal in critical frequency bands

• These effects are generally built directly into auditory models (as will be shown)
Mel Frequency Scale

Approximately linear scale until 1000 Hz
Approximately logarithmic scale from 1000 to 10000 Hz

\[ \text{Pitch in mels} = 3323 \log_{10}(1 + F/1000) \]
At 0 dB Loudness Level there is more than 60 dB variation in Intensity Level.

At 100 dB Loudness Level there is less than 10 dB variation in Intensity Level.
Critical Bandwidths

Critical Bandwidth is approximately constant between 100 and 1000 Hz. Critical Bandwidth grows logarithmically between 1000 and 10000 Hz.
Mel-Scale Filter Bank
Critical Band Shape

Trapezoidal shape is shown on a bark frequency scale – so it doesn’t look trapezoidal.
Critical Bands

4-Band filter bank that approximates logarithmic frequency spacing from 100 Hz to 3200 Hz
Auditory Masking

Strong masking signal changes the hearing threshold, and effectively masks signals that fall below the masked threshold and are within a critical bandwidth of the masking signal.
Masking can occur in time (as well as in frequency) where a strong temporal masking signal can pre-mask up to 30 msec. of signal, and post-mask up to 200 msec of signal.
What Do We Learn From Auditory Models

- Need both **short** (20 msec for phoneme-duration signals) and **long** (200 msec for syllable-duration signals) segments of speech
- **Temporal structure** of speech is important
- **Spectral structure** of sounds (formants) is important
- **Dynamic** (delta) features are important
Proposed Auditory Models

• Perceptual Linear Prediction (PLP) – H. Hermansky
• Mean-Synchrony Auditory Model – S. Seneff
• Cochlea Model – D. Lyon
• Ensemble Interval Histogram – O. Ghitza
Perceptual Linear Prediction

- Critical band (Bark) non-linear frequency resolution
  - Variable bandwidth trapezoidal integration in frequency
- Asymmetries of auditory filters
  - 25 db/Bark slope at high frequency cutoff
  - 10 db/Bark slope at low frequency cutoff
- Unequal sensitivity of human hearing versus frequency
  - Approximation to Fletcher-Munson equal loudness curve
- Intensity-loudness non-linear relationship
  - Cubic root compression (but not proper for very loud or very quiet sounds)
- Broader than critical-band integration
  - Autoregressive all-pole model, 5th order analysis => 2 resonances
Perceptual Linear Prediction

1. Compute power spectrum estimate using FFT
2. Use either a triangular window (mel-based) or a trapezoidal window (Bark-based) to integrate the power spectrum within overlapping critical bands
3. Pre-emphasize the spectrum to approximate the unequal sensitivity of human hearing versus frequency (approximates equal loudness curves)
4. Compress the spectral amplitudes with a logarithmic compressor (approximates power-law relation between intensity and loudness)
5. Perform an inverse DFT to give cepstral coefficients
6. Perform spectral smoothing (e.g., using cepstral lifter for mel-cepstral processing, or using an autoregressive model for the compressed critical band spectrum)
7. Use an orthogonal representation (e.g., convert autoregressive components back to cepstral components)
8. Perform liftering using an $n^a$ lifter
Perceptual Linear Prediction

Speech

Fast Fourier Transform

Critical-band integration and re-sampling

Equal-loudness curve

Power law of hearing

Inverse Discrete Fourier Transform

Solving of set of linear equations (Durbin)

Cepstral recursion

Cepstral coefficients of PLP model

(1) Get power spectrum

(2) Frequency axis warping (Bark scale)

(3) Convolution with critical band masking curve and down sampling

(4) Equal-loudness pre-emphasis

(5) Intensity-loudness (cubic root) amplitude warping

(6) All-pole modeling
LPC Versus PLP

LPC: pre-emphasis done at front end; PLP: uses equal-loudness filtering

LPC: linear spectral analysis; PLP: compressed critical band spectrum

LPC: cepstrally smoothed spectrum; PLP: LPC-based smoothing of spectrum
**Seneff Auditory Model**

**Mean Rate Spectrum:** firing rates of auditory nerve fibers; measure of spectral energy

**Synchrony Spectrum:** synchrony of fine temporal structure in each channel

**Lateral Synchrony (Ali):** repetition cues from adjacent channels

Stages I and II: peripheral transformations occurring in the early stages of the hearing process

Stage III: extract information relevant to perception, such as formants and enhanced sharpness of onset and offset of speech segments
**Seneff Auditory Model**

**Stage I:**
- 16 kHz sampling rate
- Pre-filter to eliminate high and low frequency components
- 40-channel critical band linear filter bank

**Stage II:**
- Hair cell synapse model
- Captures features of transformation from basilar membrane vibration to probabilistic response properties of the auditory nerve fibers

**Stage III:**
- Generalized Synchrony Detector (GSD) implements “phase locking” property of nerve fibers => enhances spectral peaks due to vocal tract resonances
- Envelope Detector (ED) computes envelope of signals from Stage II; captures very rapidly changing dynamic nature of speech; i.e., transient sounds
**Seneff Auditory Model**

**Implementation formulas**

\[ H_i(z) = HS_i(z)HP_i(z) \]

\[ HS_i(z) = \prod_{j=i}^{L} k_j (1 - X)^i (1 - Y)^i \]

\[ HP_i(z) = \left[ \frac{(1 - ZP_i z^{-1})(1 - ZP_i^* z^{-1})}{(1 - PP_i z^{-1})(1 - PP_i^* z^{-1})} \right]^2 \]

\[ y(n) = \begin{cases} G_{HW} \{1 + A \tan^{-1}[Bx(n)]\} & \text{if } x(n) > 0 \\ G_{HW} e^{A[Bx(n)]} & \text{if } x(n) \leq 0 \end{cases} \]

\[ dc(t) = \begin{cases} \mu_c [s(t) - c(t)] - \mu_b(t) & \text{if } s(t) > c(t) \\ -\mu_c c(t) & \text{if } s(t) \leq c(t) \end{cases} \]

\[ H(z) = \left( \frac{1 - \alpha_z}{1 - \alpha_z z^{-1}} \right)^{nLP} \]

\[ y(n) = \frac{x(n)}{1 + K_{AGC} x(n)} \quad H(z) = \left( 1 - \alpha z^{-1} \right) \]

\[ H_{ED}(z) = \left( \frac{1 - \alpha_{sd}}{1 - \alpha_{sd} z^{-1}} \right)^2 \quad S_{sd}(n)_i = 4 \tan^{-1} \left( \frac{1 - x_i(n) + x_i(n-M_i)}{4} \right) \]

\[ H_{SD}(z) = \left( \frac{1 - \alpha_{sd}}{1 - \alpha_{sd} z^{-1}} \right)^2 \quad i = 1, \ldots, L \]
Mean Rate and Synchrony Spectrums for word /pa’tata/

GSD spectrum has a limited number of well defined spectral lines, and tracks the dynamic modifications of the speech signal.

ED spectrum shows transitions from one phonetic segment to the next as onsets and offsets in the output representation (due to forward masking in the model).
Lyon’s Auditory Model

Models behavior of the cochlea as a non-linear, compressive, filter bank
Lyon’s Auditory Model

Cascade filter bank with 86 filters

Outer ear and middle ear provide pre-emphasis processing

Ideal Half Wave Rectification models behavior of Inner Hair Cells
Lyon’s Auditory Model

Four AGC phases in cascade

Value of gain of each stage depends on time constant from preceding output samples of adjacent channels; this reproduces masking effects

Outputs approximately represent the neural firing rates

**AGC** (passive)

<table>
<thead>
<tr>
<th>AGC target</th>
<th>τ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1°</td>
<td>0.032 640 ms</td>
</tr>
<tr>
<td>2°</td>
<td>0.016 160 ms</td>
</tr>
<tr>
<td>3°</td>
<td>0.008 40 ms</td>
</tr>
<tr>
<td>4°</td>
<td>0.004 10 ms</td>
</tr>
</tbody>
</table>

\[
a = (1 - \varepsilon) / 3 \\
b = \varepsilon / \text{target} \\
\varepsilon = 1 - e^{-1/(+ \text{time constant})}
\]
Lyon’s Auditory Model

‘Cocleogram’

‘passive’ model

Time and frequency representations of output of the auditory model

/a/, 16 kHz
Lyon’s Auditory Model

‘Correlogram’

Short time autocorrelation of each output of the auditory model

Result called correlogram

Plot shows time, frequency and autocorrelation lag

Correlogram shows where energy is located in frequency, as well as the value of the autocorrelation lag for which the signals of the cochlear channels have the same periodicity; it is possible to see how the pitch of the input signal varies in the time domain

/ɑ/, 16 kHz
Auditory Models Learnings

- **Non-uniform frequency analysis** is essential to match pitch perception.
- Some type of **logarithmic compression** is essential to match loudness perception.
- Some type of **AGC (Automatic Gain Control)** is essential to match masking perception.
- Both **temporal and spectral features** are important and information bearing cues.
- Both **long term** (~200 msec) and **short term** (~20 msec) analyses are essential to characterize syllable-length units and phoneme-length units.
Back to our problems...
Distinctive Feature Detection

1. What attributes?

2. What signal processing for each attribute?

3. What features?

4. How to optimize from training?

5. Training set label correction

Front-end Processing 1
- e.g. Auditory Model + processing

Front-end Processing 2
- e.g. Pre-emphasis

Front-end Processing M
- e.g. Pitch detector

Attribute Detector 1

Attribute Detector 2

Attribute Detector 3

Attribute Detector N

Attributes Combination
- Linear,
- MLP,
- K-L,
- etc.

Feature Detector 1
- e.g. Burst Info

Feature Detector 2
- e.g. VOT

Feature Detector L

Training Set

Training Set Phonetic Feature Labels
Q1. What Attributes?

- Conventional Wisdom Attributes

- Mel-frequency (or Bark frequency) direct or auditory model-based **cepstral coefficients** (~13) from short analysis frames (~20 msec)

- **Delta mfcc** using order of 5 frames (~100 msec)

- **Delta-delta mfcc** using order of 3 frames (~60 msec)

- **Gross speech attributes** including voiced/unvoiced/silence, pitch, SNR estimates

- **Temporal speech attributes** including VOT (when appropriate), burst duration (when appropriate), burst frequency, syllable duration, sound duration

- **Derived measures** such as spectral flatness, relative amplitudes, ratio of frequency band energies, etc.
### Q1. - Desired Speech Attributes

<table>
<thead>
<tr>
<th>Spectral</th>
<th>Temporal</th>
<th>Short Term (20 ms)</th>
<th>Long Term (200 ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mfcc</td>
<td>voiced/unvoiced/silence</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>spectral flatness</td>
<td>pitch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>relative band energies</td>
<td>SNR</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VOT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>burst duration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>unvoiced duration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>syllable duration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>delta(mfcc)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>delta-delta(mfcc)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q2. What Front-End Processing?

- Standard processing (LPC, FFT)
- Simple pre-emphasis
- Pitch detection, Voiced-Unvoiced-Silence detection
- Auditory Model Processing (PLP, MFCC, Seneff, Lyon)
- Other processing?
Q3. What Features? - Desired Speech Features

- Initial list of twelve pairs of binary distinctive features (Jakobson, Fant and Halle)
  - 1. Vocalic/non-vocalic
  - 2. Consonantal/non-consonantal
  - 3. Interrupted/continuant
  - 4. Checked/unchecked
  - 5. Strident/mellow
  - 6. Voiced/unvoiced
  - 7. Compact/diffuse
  - 8. Grave/acute
  - 9. Flat/plain
  - 10. Sharp/plain
  - 11. Tense/lax
  - 12. Nasal/oral
- These are binary valued speech production features; they are not efficient (2**12=4096, but there are significantly fewer phones and phonemes)
- English is characterized by 9 pairs of these features (still not efficient 2**9=512)
- 61 phones in TIMIT representation; 39 phonemes plus silence model used in most modern speech recognition systems

Conventional phonetic feature sets are not efficient and require measurements that are difficult to make reliably
Q3. Distinctive Features

Classify non-vowel/non-diphthong sounds in terms of distinctive features

- place of articulation
  - Bilabial (lips)—p,b,m,w
  - Labiodental (between lips and front of teeth)-f,v
  - Dental (teeth)-th,dh
  - Alveolar (front of palate)-t,d,s,z,n,l
  - Palatal (middle of palate)-sh,zh,r
  - Velar (at velum)-k,g,ng
  - Pharyngeal (at end of pharynx)-h

- manner of articulation
  - Glide—smooth motion-w,l,r
  - Nasal—lowered velum-m,n,ng
  - Stop—constricted vocal tract-p,t,k,b,d,g
  - Fricative—turbulent source-f,th,s,sh,v,dh,z,zh,h
  - Voicing—voiced source-b,d,g,v,dh,z,zh,m,n,ng,w,l,r
  - Mixed source—both voicing and unvoiced-j,ch
  - Whispered--h
Q3. Distinctive Phoneme Features

<table>
<thead>
<tr>
<th>Place</th>
<th>p</th>
<th>k</th>
<th>t</th>
<th>b</th>
<th>d</th>
<th>g</th>
<th>f</th>
<th>thin</th>
<th>s</th>
<th>sh</th>
<th>v</th>
<th>the</th>
<th>z</th>
<th>azure</th>
<th>m</th>
<th>n</th>
<th>ng</th>
<th>l</th>
<th>r</th>
<th>w</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>bilabial</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>labiodental</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>dental</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>alveolar</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>palatal</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>velar</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>pharyngeal</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

**Manner**

| glide      | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| nasal      | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | + | + | - | - | - |
| stop       | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | - | - | - | - | - |
| fricative  | - | - | - | - | + | + | + | + | + | + | + | + | + | + | + | + | - | - | - | - | - |
| voicing    | - | - | - | - | + | + | - | - | + | + | + | + | + | + | + | + | + | + | + | + | + |

**FIGURE 17.7** Binary distinctive feature set of Jakobson et al. From [10].

- the brain recognizes sounds by doing a distinctive feature analysis from the information going to the brain
- the distinctive features are somewhat insensitive to noise, background, reverberation => they are robust and reliable
Q3. – Binary Valued Features vs. Multi Valued Features

- Binary valued features vs. multi valued features – Stephenson’s result (1988). Use 12 MFCCs plus energy as input.

- Binary valued:
  - MLP with 250 Hidden units
- Multi valued:
  - 20 ~ 80 hidden layers for each feature
- Each feature plus silence

1. ANN more reliable than HMM in measuring these modified features
2. 8 multi-valued features more efficient than conventional binary feature set
Q4. How to Optimize Attributes Combination Methods

- Linear Interpolation
- Multi-Layer Perceptron (MLP)
- Karhunen-Loeve Expansion (KL)
- Principal Component Analysis (PCA)
- Support Vector Machines (SVM)
- Decision Trees
- K-Means Algorithm
- Other methods?
Q4. Optimize Combination from Training

• Recursive training, adjusting weights between iterations
• Convergence issues:
  – Convergence to local maxima
  – Recognition of global maxima (when achieved)
• What criteria to use to calculate the effectiveness of the combining methods
  – Expectation Maximization (EM)
  – Mutual Information (MI)
  – Maximum Entropy (ME)
  – Other methods?
• Initialization issues:
  – Random starts
  – Pre-trained or hand selected starts
Q5. Training Set Label Correction

- TIMIT labels assume perfect production rules according to simple rules of language.
- One phoneme may have many labels, but only a single one is chosen for TIMIT transcriptions, e.g., voiced consonants can become devoiced at final position (/z/ => /s/ at end of sentence) but either /z/ or /s/ is chosen, not both.
- Imprecise segmentation in TIMIT – due to human decisions and uncertainty in boundary locations.
- Re-test, investigate source of gross errors, re-label the training set (when appropriate and reasonable).
### Q5. - Training Set Imbalance

<table>
<thead>
<tr>
<th>Centrality</th>
<th>Frontback</th>
<th>Manner</th>
<th>Phonation</th>
<th>Place</th>
<th>Roundness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central (6)</td>
<td>Back (23)</td>
<td>Vowel (37)</td>
<td>Voiced (61)</td>
<td>Low (7)</td>
<td>Unrounded (79)</td>
</tr>
<tr>
<td>Full (34)</td>
<td>Front (62)</td>
<td>Fricative (18)</td>
<td>Unvoiced (27)</td>
<td>Medium (12)</td>
<td>Rounded (9)</td>
</tr>
<tr>
<td>Nil (49)</td>
<td>Nil (3)</td>
<td>Approximant (9)</td>
<td>Silence (12)</td>
<td>High (8)</td>
<td></td>
</tr>
<tr>
<td>Silence (12)</td>
<td>Silence (12)</td>
<td>Nasal (18)</td>
<td></td>
<td>Labial (8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Occlusive (18)</td>
<td></td>
<td>Dental (5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Silence (12)</td>
<td></td>
<td>Coronal (26)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Velar (10)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Glottal (3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Silence (12)</td>
<td></td>
</tr>
</tbody>
</table>

Percentage of features in the TIMIT training set for different phonetic features

(From T. Stephenson’s result. 1988.)
Front End Auditory Processing -- Ali

- Auditory effects used:
  - Bark-scaled filter bank of 36 filters, 0.5 Bark apart
  - Compressive non-linearity
  - Half-wave rectification
  - Automatic gain control (AGC)
  - Short term adaptation
  - Forward masking
- 2 outputs: mean-rate, synchrony
Front-end Auditory Processing -- Ali

- ALSD as a modification of GSD

\[ ALSD_i = \frac{1}{n} \sum_{k=i-n1}^{i+n2} GSD_i(y_k) \]

- Decrease individual harmonics and spurious peaks
- Recognize 4 vowels: /ae/, /iy/, /aa/ and /uw/, accuracy is approximately 81% for clean speech, 79% for SNR=10dB. Approximately 14% higher than GSD methods
Fricative Detection Features -- Ali

- Seneff auditory model used for front end processing
- Duration of Unvoiced Portion (DUP)
  - Need V/U detector
  - Need DUP threshold (<60 msec => voiced, >100 msec unvoiced)
- Relative Amplitude (RA) – discriminates place, large RA => s,sh,z,zh; small RA => f,th,v,dh
  - Using mean rate outputs
  - Normalized with respect to nearest vowel
- Spectral Flatness
  - Maximum normalized spectral slope (MNSS)
- Spectral Shape and Peak Location
  - Spectral shape used to determine place of articulation
  - MDP-Most Dominant Peak location (from GSD output); separates sh,zh from other fricatives
  - SCG-Spectral Center of Gravity
- Place of Articulation Detection
  - Combined decision using MNSS, SCG, MDP
Fricative Detection Features -- Ali

Compare histograms of discrimination abilities of RA (left) and MNSS (right)
Fricative Detection Features -- Ali

- Overall classification results: 87%
- If compare RA and MNSS results only:
  - Overall: 90% -> 94%
  - /s, z/: 95% -> 99%
  - /f, v, th, dh/: 78% -> 91%
  - /sh, zh/: 98% -> 85%
  - MNSS discriminate much better /f, v, th, dh/ than RA
  - But /sh, zh/ results decrease from 98% to 85%. Deduction: there must be some information conflict between RA and MNSS for palatals /sh/ and /zh/!
- Can we find other features that resolve this conflict, and get better results for all fricatives?
Stop Consonant Detection -- Ali

- Voicing, place of articulation estimated
- Feature-based decision tree algorithm
- Voicing Detection
  - Voicing during closure (pre-voicing) => voiced stop
  - VOT (Voice Onset Time) => discriminates between V and UV (larger for unvoiced stops)
  - Closure duration – cue to flag which VOT threshold to use
  - Only works for initial and medial stops
- Place of Articulation Detection
  - Burst frequency—BF
  - Second formant of following vowel
  - Maximum normalized spectral slope (MNSS)
  - Burst frequency prominence (most prominent peak in synchrony output during stop release => largest amplitude or largest spectral slope)
  - Formant transition before and after stop
  - Voicing decision
- Synchrony output shows formants and dominant peaks, and is relatively insensitive to noise
Stop Consonant Detection -- Ali

• The burst frequency BF was the most important feature in the place of articulation detection. But BF is highly dependent on the next vowel, detected using F2 of the neighboring vowel represented by ALSD.

• But only valid for initial and medial positions. Stops in absolute final positions and stops followed by fricatives don’t follow the pre-voicing, VOT and closure rules. What rules should be used in the latter two conditions?
General Phoneme Detection -- Ali

- Total energy
- Spectral center of gravity (SCG)
- Duration
- Low, medium and high frequency energy
- Formant transitions
- Silence detection
- Voicing detection
- Rate of change of energy in various frequency bands
- Rate of change of SCG
- Most prominent peak frequency
- Rate of change of most prominent peak frequency
- Zero crossing rate
Ali Feature Detector

- Input speech
  - Front-end Processing
    - Preliminary Segmentation And Categorization
      - Silences
      - Obstruents
      - Sonorants
        - Obstruent Segmentation
          - Stops
            - Voicing Detection
              - Place of Articulation Detection
        - Fricatives
          - Voicing Detection
            - Place of Articulation Detection
        - Fine Sonorant Identification
          - Segmentation within a Sonorant Segment

- Recognized Phoneme Certainty Factors
Ali Feature Detector

• Discussion
  – Good features selected and excellent performance
  – Need new features to distinguish inside certain groups
  – May find more information from the error patterns, and new features to further refine the results, e.g. get less information conflict
  – May need better criteria to calculate the information contained in each feature