AUDITORY MODELS AS FRONT-ENDS FOR SPEECH RECOGNITION

Fernando S. Perdigão, Luís V. Sá

Instituto de Telecomunicações - Pólo de Coimbra
Dept. Eng. Electrotécnica - Pólo II da Univ. Coimbra
3030 COIMBRA - PORTUGAL
fp@it.uc.pt
MOTIVATION:

• Auditory Models have shown to be superior in recognition tasks when environment degrades (additive noise + linear filtering).
• There is not a deep understanding of their functioning.

OBJECTIVE:

• Verify if there are advantages in the characterization of speech signals using models of the auditory periphery.
• Compare Auditory Models operation with other signal processing schemes.
MAIN CHARACTERISTICS OF PERIPHERAL AUDITORY PROCESSING

COCHLEA

HAIR-CELLS

FIBERS

Mean-Rate Representation

$y_1(n)$

$y_N(n)$

$x(n)$

$H_N$

$H_I$

$log$ freq.

$dB$

level

time

LPF

$f_1(n)$

$\ldots$

$\ldots$

$\ldots$

Half-Wave Rectification

(w/ compression)

Adaptation

Loss of Synchrony
FILTER-BANK RESPONSES

a) Cochlear Model  
b) Seneff - Stage I  
c) Gamma-Tone

Frequency responses of filter banks with 35 channels. $f_s=8$kHz. 
The responses shown corresponds to channels 4, 11, 18, 25 and 32 (taps) or CFs of 3, 1.8, 1.1, 0.6 and 0.3 kHz (Gamma-Tone).
Block Diagrams of IHC/Synapse Models

Meddis’ Model

\[ y_n \xrightarrow{A} \text{HWR} \xrightarrow{\frac{g_z n}{z_n + B}} \text{H}_1 \xrightarrow{-q_n} \text{H}_2 \xrightarrow{f_n} \]

\[ H_1(z) = \frac{1}{s + l_0 + r_0} \left|_{z=e^{(-1)f_i}} \right. \]

\[ H_2(z) = \frac{s^2 + s(l_0 + r_0 + x_0) + l_0 x_0}{(s + y_0)(s + x_0)} \left|_{z=e^{(-1)f_i}} \right. \]

Seneff’s Model

\[ y_n \xrightarrow{\text{NL}} \text{HWR} \xrightarrow{z_n} \text{H}_1 \xrightarrow{\nu_n} \text{AGC} \xrightarrow{f_n} \]

\[ H_4(z) = \left( \frac{1 - \alpha z^{-1}}{1 - \alpha z^{-1}} \right)^4 \]

\[ f_n = \frac{\nu_n}{1 + K_{\text{AGC}} q_n} \]

Martens-Immerseel’s Model

\[ y_n \xrightarrow{A} \text{HWR} \xrightarrow{\nu_n} \text{AGC} \xrightarrow{f_n} \]

\[ H_2(z) = r \frac{1 - a_1}{1 - a_1 z^{-1}} + (1 - r) \frac{1 - a_2}{1 - a_2 z^{-1}} \]

\[ f_n = \frac{\nu_n f_{sw}}{B + q_n z} \]

HWR: Half-Wave Rectification. AGC: Automatic Gain Control (as a function of \( q_n \)).
NL represents the Seneff’s model nonlinearity: \( 1 + \arctan(B y_n) \) for \( y_n > 0 \) and \( \exp(ABy_n) \) for \( y_n < 0 \).
The blocks \( H_k \) represent kth order filters.
INNER-HAIR CELLS AND NERVE FIBERS

Main Characteristics:
- Half-Wave Rectification (IHC transduction is directional)
- Auditory Fibers Firing-Rate
  - Spontaneous and Saturation values
  - Threshold of excitation
  - Adaptation
  - Synchrony

IHC transduction

Time-Histogram of Discharges
MODEL RESPONSES
(High Spontaneous-Rate Fiber)

RATE-INTENSITY CURVE
(steady excitation)

ADAPTATION

Threshold
input level

input burst

1 ms average

Martens  Seneff

Martens  Meddis

sontaneous rate

saturation

input level [dB]

Threshold

0 20 40 60 80 100

Martens  Meddis

Seneff

0 40 80 120 160

0 2 4 6 8 10

0 1000

f_{n}

0 100 200 300 400 500 600

0 0.05 0.1 0.15 0.2 0.25
Problems with IHC/Synapse Models:

- Due to non-linearities they have to be simulated in time, on a sample-by-sample basis. This demands a high computational load.
- The output mean-rate is then decimated to have a feature representation every 10ms.

But, the study carried out shows that:

*Adaptation can be reasonably modeled in terms of the short-term envelope of the energy (or RMS value) of the sub-band signals.*

⇒ *Functional Model of Adaptation*
  
  (proposed model)
  
  ⇒ Energies are computed in frequency domain (using the FFT).
  
  ⇒ Adaptation is modeled with RMS values.
Functional Model of Adaptation

\[ x[n] \rightarrow \text{Hamming} \rightarrow \text{FFT} \rightarrow H_1 \rightarrow \ldots \rightarrow \text{RMS} \rightarrow Y_i \rightarrow V_i \rightarrow f_{\text{rate}}V_i \rightarrow F_i \]

\[ Y[m] = \frac{1}{N} \sum_{k=0}^{N-1} |X_m(k)H_i(k)|^2 \]

\( H_i \): filter \( i \) of the filter bank. \( F_i \): output firing rate. 
\( A \): rate threshold

Adaptation Model: Martens-Immerseel model with time constants of 20 and 40ms, operating with a sampling frequency of 100Hz (frame-rate).
NOISE SUPPRESSION (with Martens’ Model)

• Mean-Rate representation is very sensitive to noise (because noise adapts the responses).

• [Vereecken & Martens, Eurospeech’95] propose a Center-Clipper in front of the IHC model. The clipping level is chosen in order to keep the output of HWR almost constant. The clipping level depends on the noise variance and pdf (Gaussian assumed).

\[
y_n \rightarrow \text{Center-Clipper} \rightarrow HWR \rightarrow \text{AGC} \rightarrow f_n
\]

• An empirical study showed that RMS values, \( Y_i[m] \), present a gamma distribution. An analogous noise suppression technique was used, based on:

\[
V_i[m] = A + \max(Y_i[m] - \Delta, 0)
\]
GAMMA DISTRIBUTION FUNCTION

\[ f_x(x) = \frac{c^{b+1} x^b e^{-cx}}{\Gamma(b+1)} u(x) \]

\[ \mu_x = \frac{b+1}{c} \quad , \quad \sigma_x^2 = \frac{b+1}{c^2} = \frac{\mu_x}{c} \]

Fitting of \( Y_i[m] \) pdf:

\[ Y_i[m] \quad \sigma=100, \quad i=1 \]
\[ b_i=38.6 \quad c_i=2.01 \]

\[ Y_i[m] \quad \sigma=100, \quad i=17 \]
\[ b_i=16.0 \quad c_i=1.43 \]

\[ Y_i[m] \quad \sigma=100, \quad i=35 \]
\[ b_i=5.77 \quad c_i=1.08 \]
Characterization of Adaptation as a Filtering Operation (with tone-bursts)

Model to characterize adaptation as a linear filtering in the case of sinusoidal bursts. $V_M$ represents the sinusoid amplitude that adapts the response to a burst of amplitude $V_P$.

Filter responses of filter $h[n]$ according to top figure for several states of adaptation.
In the fiber dynamic range $\rightarrow$ almost log compression.
Fiber Threshold $\rightarrow A \leftrightarrow 1/J$ in J-RASTA framework
Adaptation $\rightarrow$ Rasta filtering
ISOLATED-DIGIT EXPERIMENTS

Database: Telephone-speech digits, \( \approx \)800 speakers, 4200 digits (2100 for training).
Recognizer: CDHMM, 7 states, mixture with 6 components, diagonal covariance matrices. Silence model.
Pre-Processing: signal normalization (n)
Post-Processing: clipping (c) or spectral subtraction (s)
Distortion: adding white noise (SNR=20dB) and/or filtering the speech signals

<table>
<thead>
<tr>
<th>Functional Model</th>
<th>Label</th>
<th>Clean</th>
<th>SNR=20dB</th>
<th>Filter</th>
<th>Noise+Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-Rasta</td>
<td>97.57</td>
<td>90.22</td>
<td>97.55</td>
<td>89.66</td>
<td></td>
</tr>
<tr>
<td>J-Rasta</td>
<td>96.27</td>
<td>89.59</td>
<td>93.71</td>
<td>88.56</td>
<td></td>
</tr>
<tr>
<td>MFCC_E_D</td>
<td>97.67</td>
<td>80.49</td>
<td>94.81</td>
<td>47.57</td>
<td></td>
</tr>
<tr>
<td>LPC_E_D</td>
<td>97.95</td>
<td>85.90</td>
<td>94.96</td>
<td>62.77</td>
<td></td>
</tr>
<tr>
<td>GMedH</td>
<td>94.00</td>
<td>63.33</td>
<td>88.96</td>
<td>38.16</td>
<td></td>
</tr>
<tr>
<td>SSenMn</td>
<td>88.01</td>
<td>68.00</td>
<td>87.30</td>
<td>54.25</td>
<td></td>
</tr>
<tr>
<td>HMarHc</td>
<td>92.84</td>
<td>84.43</td>
<td>88.88</td>
<td>76.64</td>
<td></td>
</tr>
<tr>
<td>HMarHnc</td>
<td>96.95</td>
<td>86.93</td>
<td>94.22</td>
<td>87.13</td>
<td></td>
</tr>
<tr>
<td>GMarHnc</td>
<td>97.85</td>
<td>90.76</td>
<td>96.79</td>
<td>93.01</td>
<td></td>
</tr>
<tr>
<td>GFMAns</td>
<td>97.19</td>
<td>94.45</td>
<td>93.77</td>
<td>91.73</td>
<td></td>
</tr>
</tbody>
</table>

G: Gamma-tone filter-bank
H: Cochlear model derived filter-bank
S: Seneff’s filter-bank
CONCLUSIONS

• Auditory Models globally outperforms standard representations for this recognition task.

• The Functional Model of Adaptation works as well as or better than models operating in the time domain. It is almost as efficient as in MFCC framework!

• There is a close relationship between Mean-Rate and J-RASTA representations.

• Mean-Rate representation is very sensitive to noise and level variation in signals. Normalization and noise compensation is needed.