Machine recognition of speech trained on data from New Jersey Labs

- tested in New Jersey: 2.8% error
- tested in Colorado: 60.7% error

- Frequency response (peak around 5 Hz)
- Impulse response (effective length around 200 ms)

- Sensitivity of hearing to modulation peaks at about 4 Hz
  - Riesz 1928, Zwicker 1952, ... 
- Modulation transfer function of primary auditory cortex peaks at about 4 Hz
  - Schreiner via Greenberg, personal communication 1997
- Modulation spectrum of speech peaks at about 4 Hz
  - Houtgast and Steeneken 1976
- Intelligibility of speech significantly impaired when 4 Hz modulation frequency component attenuated

- Frequency discrimination of short stimuli improves up to about 200 ms
- Loudness of equal-energy stimuli grows up to about 200 ms
- Minimum detectable silent interval indicates time constant of about 200 ms
- Effect of forward masking lasts about 200 ms
RASTA filter

![Graph of attenuation vs modulation frequency]

Perception of modulations
(Riesz 1923)

![Graph showing perception of modulations]
1-12 Hz Passband

• sensitivity of hearing to modulation peaks at about 4 Hz
  – Riesz 1928, Zwicker 1952, ...

• modulation transfer function of primary auditory cortex peaks at about 4 Hz
  – Schreiner via Greenberg, personal communication 1997

• modulation spectrum of speech peaks at about 4 Hz
  – Houtgast and Steeneken 1978

• intelligibility of speech significantly impaired when 4 Hz modulation frequency component attenuated

Relative importance of various components of modulation spectrum of speech for speech intelligibility and for ASR
RASTA filter

- average four neighboring frames
- subtract exponentially decaying past values ($\tau=170$ ms)

Masking in Time

- suggests ~250 ms buffer (critical interval) in auditory system
  - what happens outside the critical interval, does not affect detection of signal within the critical interval
~ 200 ms length of impulse response

- frequency discrimination of short stimuli improves up to about 200 ms
- loudness of equal-energy stimuli grows up to about 200 ms
- minimum detectable silent interval indicates time constant of about 200 ms
- effect of forward masking lasts about 200 ms

syllable-length buffer of human hearing?
Formant-Less Vowel

original speech

filtered speech

spectrogram from RASTA
Data Do Not Lie

Prof. Frederick Jelinek: “Airplanes don’t flap their wings”.

S. Lohr, New York Times, March 6, 2011

“Airplanes do not flap wings but have wings nevertheless,…..

Of course, we should try to incorporate the knowledge that we have of hearing, speech production, etc., into our systems,…. 

F. Jelinek, Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan, Speech Communication 18, 1996. 242–293

Linear Discriminant Analysis (LDA)

Linear discriminants: eigenvectors of $S^{-1}_WS_B$

$S_W$ - within-class covariance matrix
$S_B$ - between class covariance matrix

- Needs labeled data
- Within-class distributions assumed Gaussian with equal $\sigma$ (take log of power spectrum)
Spectral Basis from LDA

LDA gives basis for projection of spectral space

Psychophysics:
Critical bands of human hearing are broader at higher frequencies

Physiology:
Position of maximum of traveling wave on basilar membrane is proportional to logarithm of frequency
Sensitivity to Spectral Change
(Malayath 1999)

Two ways of using LDA

LDA gives basis for projection of spectral space

LDA gives FIR filters for processing time trajectories of spectral energies
RASTA Filters from Speech Data

frequency responses
(1st discriminant in all frequency channels)

impulse responses

original RASTA filter

engineering

effect (signal)

cognitive signal

effect (signal)

perception

good engineering could be consistent with biology

- physiology of sensory organs
- psychophysics of perception
- emulation of the knowledge in engineering
C. Shannon: Communication in Presence of Noise

combination of channel and signal spectrum should be as flat (as random-like) as possible

Forces of Nature

if signal could be controlled (e.g. speech)
  – put more signal where there is less noise
  – sensory signal optimized for a given communication channel
Evaluate spectra within a given speech sound relative to neighboring sounds

Mutual Information Between Phoneme Labels and Measurement(s) in Time

H. Yang et al 2000, F. Li (unpublished)

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$I(X;Y|Z) = \sum_{z \in Z} \sum_{x \in X} \sum_{y \in Y} p(x,y,z) \log \frac{p(y|x,z)p(x,z)}{p(y|x)p(x|z)}$$
Auditory cortical receptive fields
from N. Mesgarani

Time-frequency patterns that optimally excite a given cortical neuron

Most often frequency-localized and often rather long

1st principal component along temporal axis from about 300 STRFs
Nima Mesgarani (in preparation)

Short Term Spectral Envelope?

Ear is frequency selective!

Simultaneous masking: Sound elements outside a critical band do not corrupt decoding of elements inside the band

Temporal masking: Sound elements outside a critical interval (about 250 ms) do not corrupt decoding of elements inside the interval

\[ P(\varepsilon) = \prod P(\varepsilon) \]

Human listeners recognize speech in independent frequency bands

Human listeners recognize speech in independent frequency bands

Fletcher et al at the 1993 Summer Workshop at Rutgers University

To recognize phoneme one needs to collect information distributed over the whole syllable

Kozhevnikov and Chistovich (Speech: Articulation and Perception, 1965)
Ear is frequency selective

However, it is NOT to derive spectrum of the signal but to yield frequency-localized temporal patterns, which carry the information about underlying acoustic events.
Away from Short-Term Spectrum

ΔT

\[ s(t_0) \]

spectrum of the short segment

dr to human hearing

Frequency Domain Linear Prediction (FDLP)

FDLP
- means for all-pole estimation of Hilbert envelopes (instantaneous spectral energies) in individual frequency channels

FDLP • means for all-pole estimation of Hilbert envelopes (instantaneous spectral energies) in individual frequency channels