Commercializing Auditory Neuroscience

Lloyd Watts
Audience, Inc.

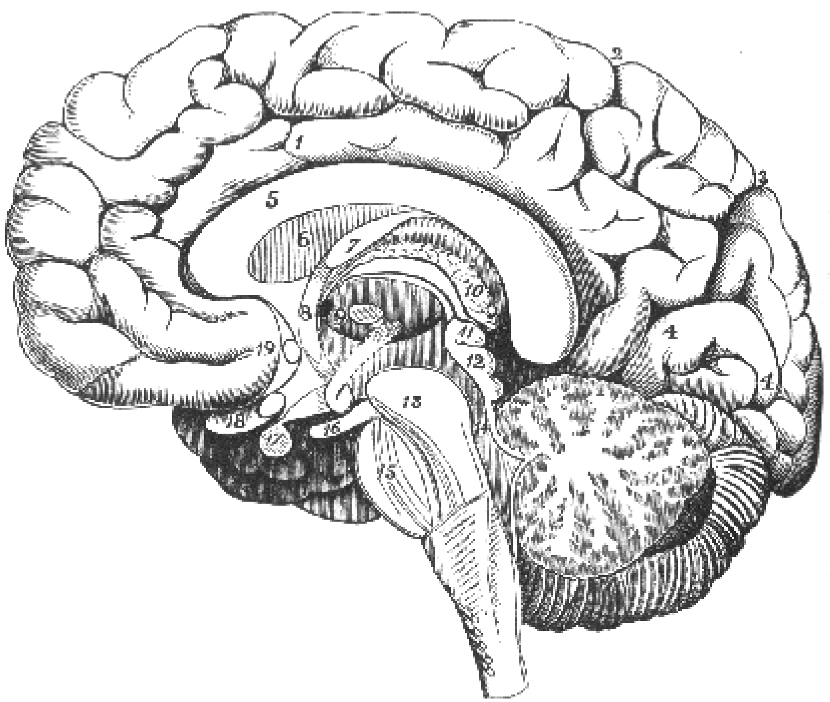
U.S. Frontiers of Engineering Symposium
Dearborn, Michigan
September 21-23, 2006

AUDIENCE™

Overview

- Can we build a machine that hears the way human beings do?
  - Original passion: music transcription
  - Reverse-engineer the auditory pathway, based on neuroscience
  - Do we know enough about the brain? Are computers capable enough?
- If so, can we build a commercially successful company out of it?
  - Can we raise the money (i.e., convince the investors)?
  - What application to shoot for?
    - Music Transcription? No…
    - Speech recognition? No…
    - Noise suppression for Cell-phones? Yes!
  - Building a team, really executing
  - Is it a chip company, or a software company?
  - Investors, Customers, Employees, Advisors all have to see short-term progress, and long-term return

Do we know enough about the brain to build one?



(Gray’s Anatomy, 1901) (Moravec, 1998)

(Kozyrakis et al., 2001)

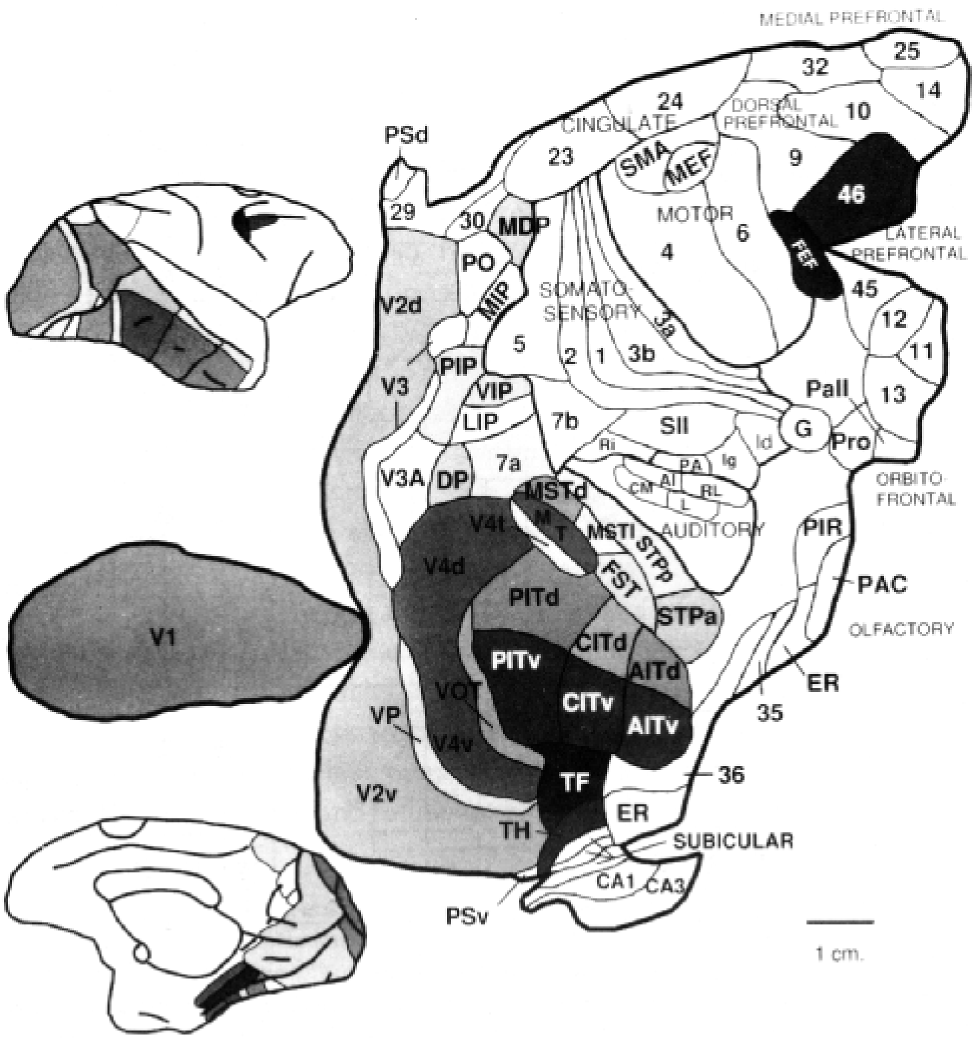
- ~10^{11} neurons, ~10^{14} synapses
- ~10^{16} Ops/s, ~ 10^{14} MByte
- ~20 W power consumption
- ~10^{6} GOPs/W efficiency (compared with ~3 GOPs/W for current HW)
- V_{dd}(brain)=80mV accounts for nearly 3 orders of magnitude of power efficiency, trades power consumption against area/cost/yield $P=CV^2f$
- Vast variety of cell types
- Thousands of modules/regions

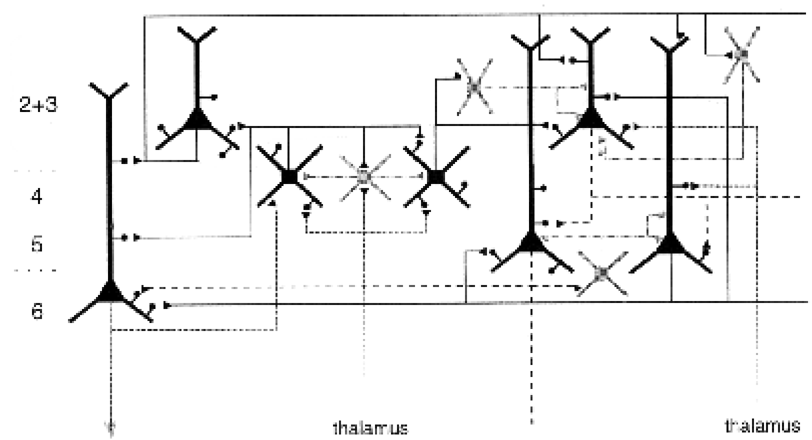
being studied by

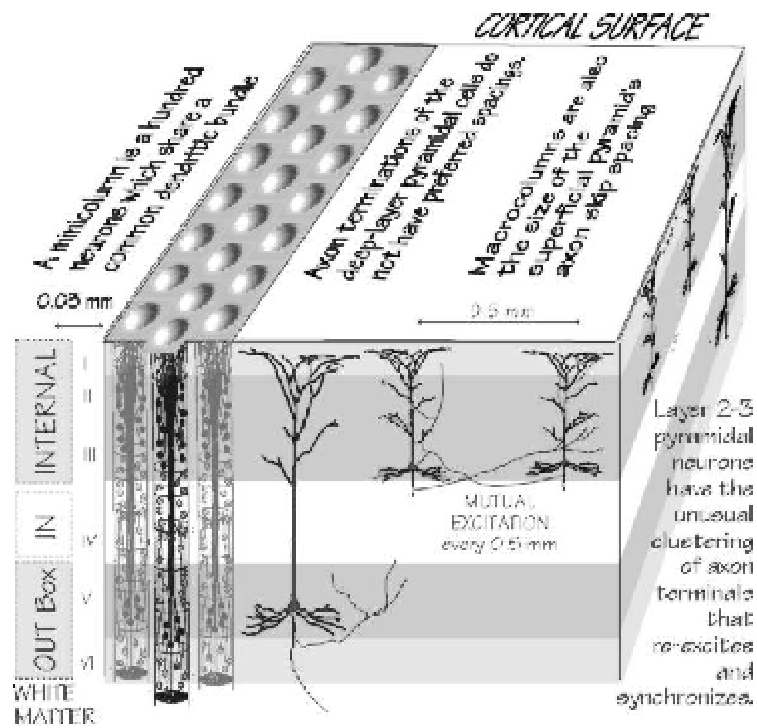
- ~30,000 neuroscientists
- **No individual knows everything.**

AUDIENCE

Dramatic increase in knowledge in last 20 years…



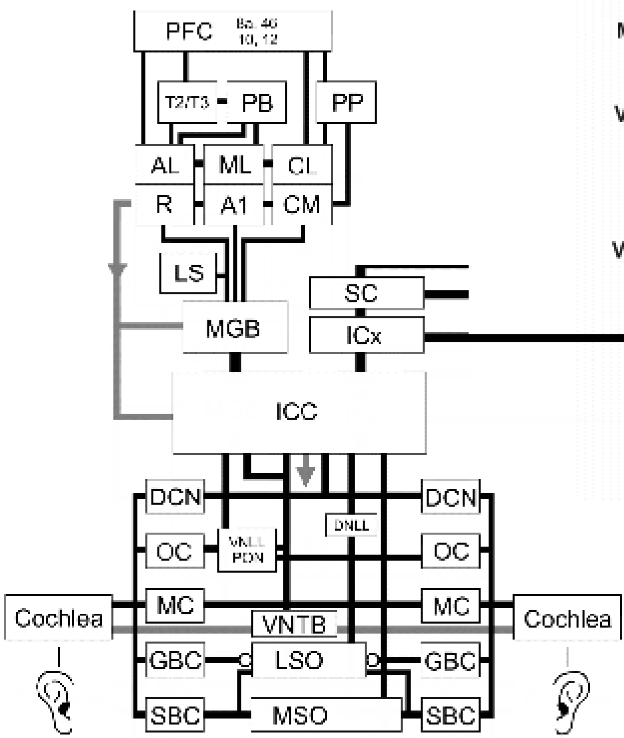




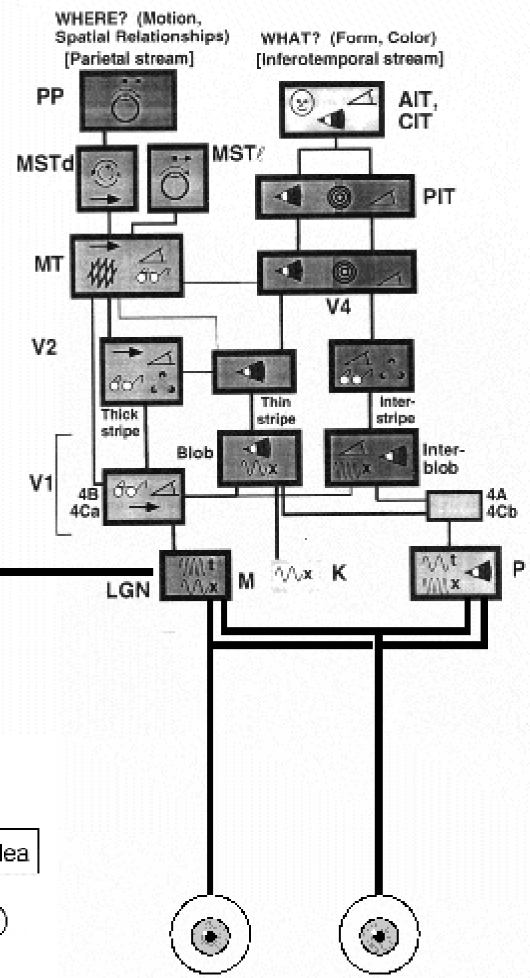
(van Essen and Anderson, 1990)
(Douglas and Martin, 1998), (Calvin, 1996)

Collectively, do we know enough about the brain to *begin* building a realistic model?

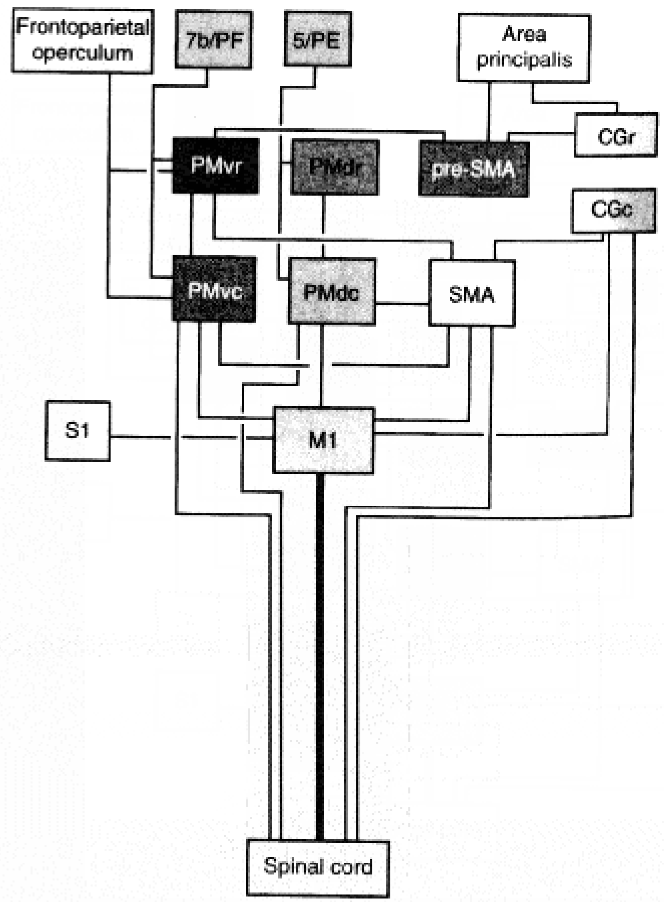
Auditory



Visual



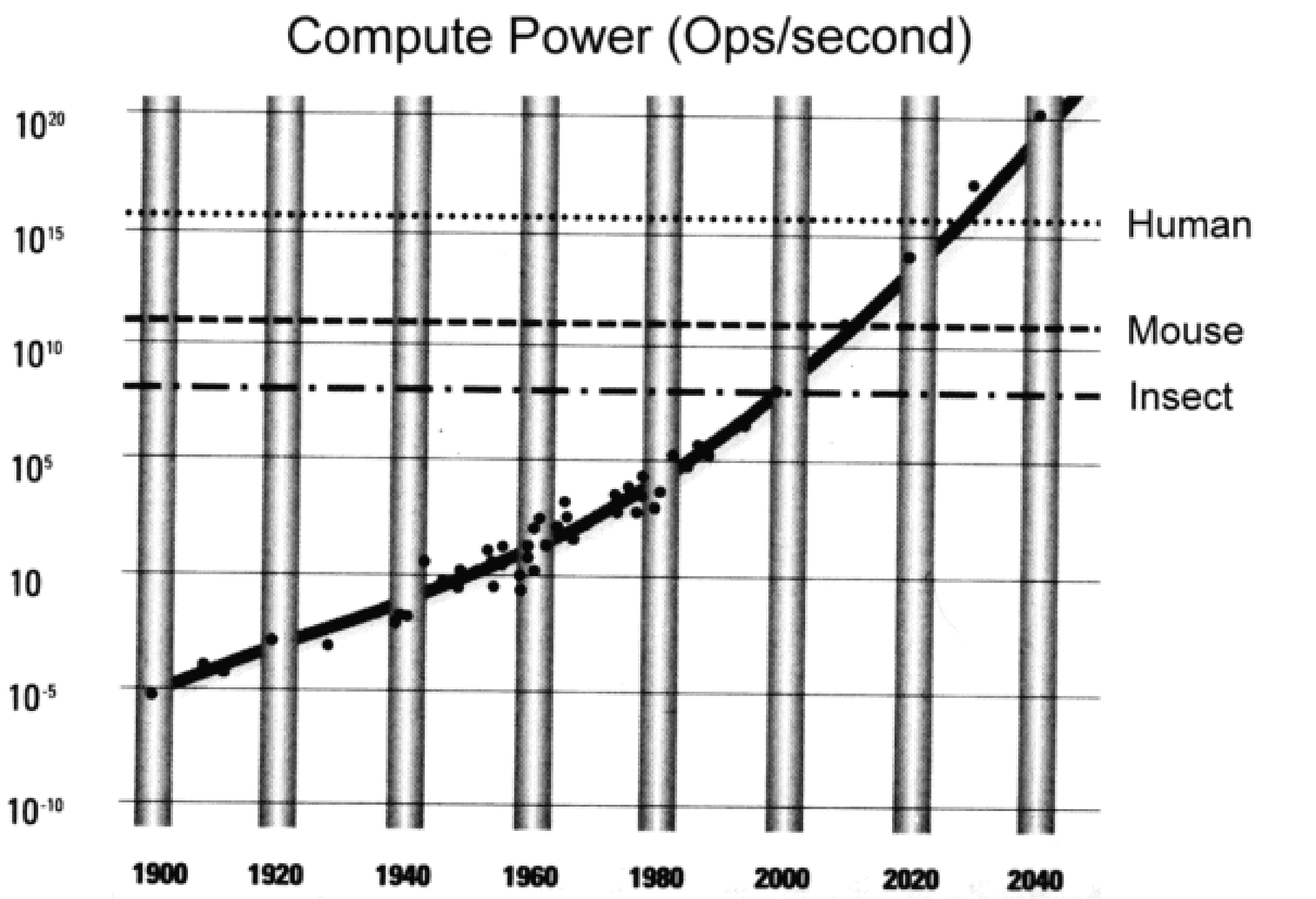
Motor



etc...

(Kiang, Oertel, Covey, Rauschecker) (van Essen and Gallant, 1994) (Zigmond et al., 1999)

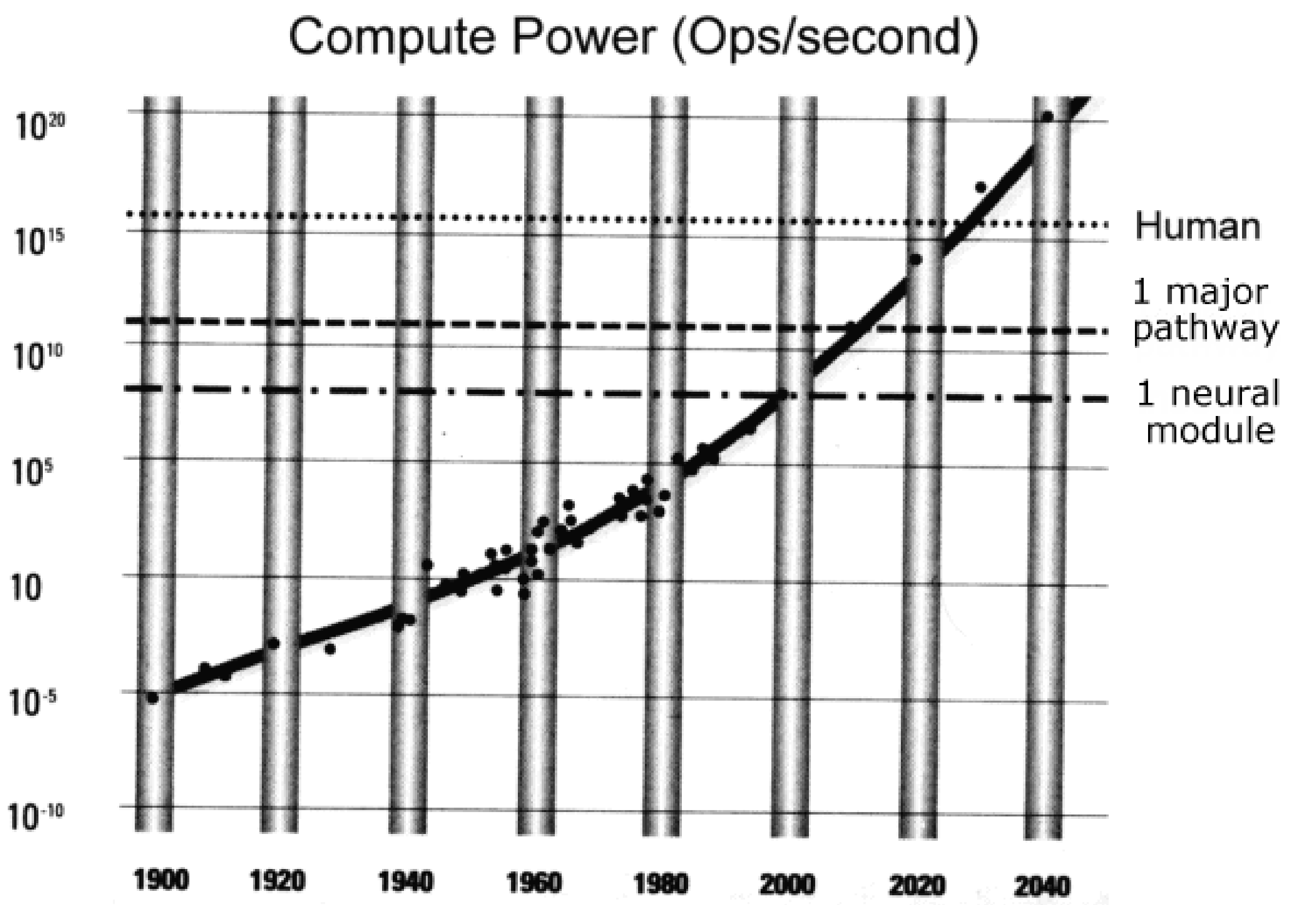
When will computers be capable enough?



(Ray Kurzweil, *The Age of Spiritual Machines*, 1999)

AUDIENCE

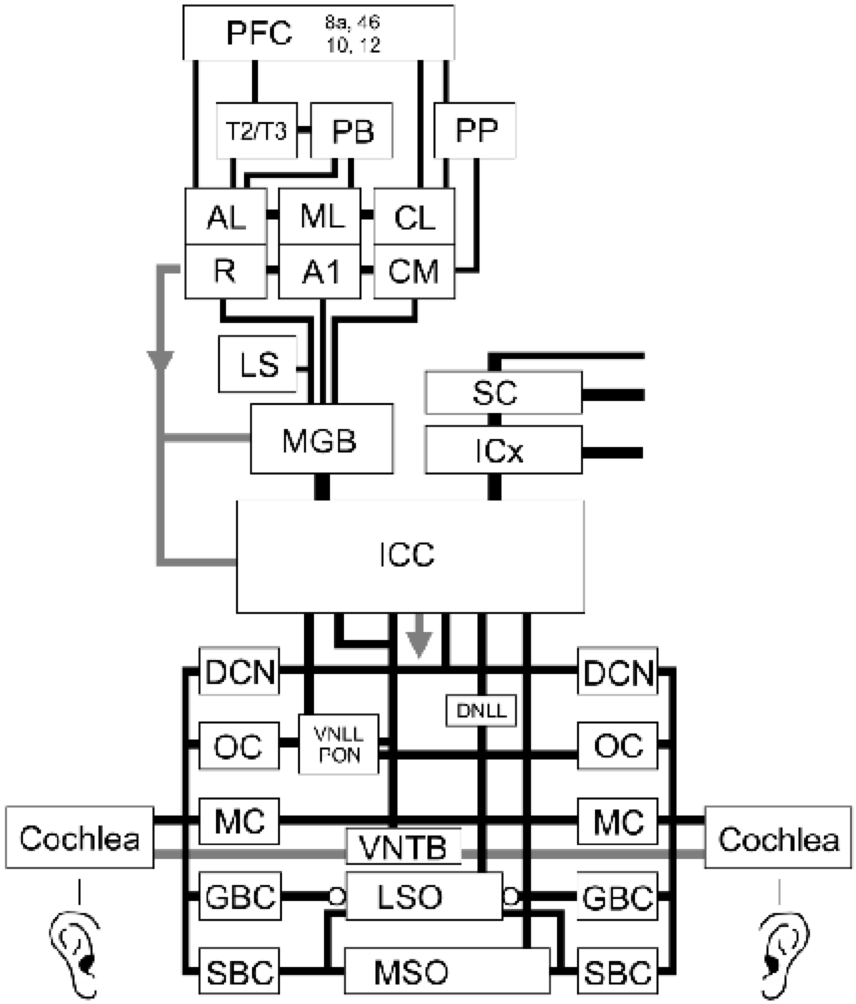
When will computers be capable enough?



Are there fundable, commercially viable applications for the major pathways (audition, vision, motor)?

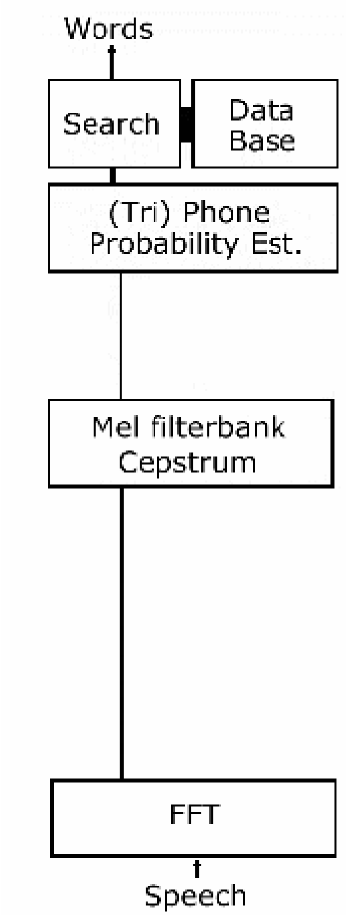
AUDIENCE

Auditory Pathway

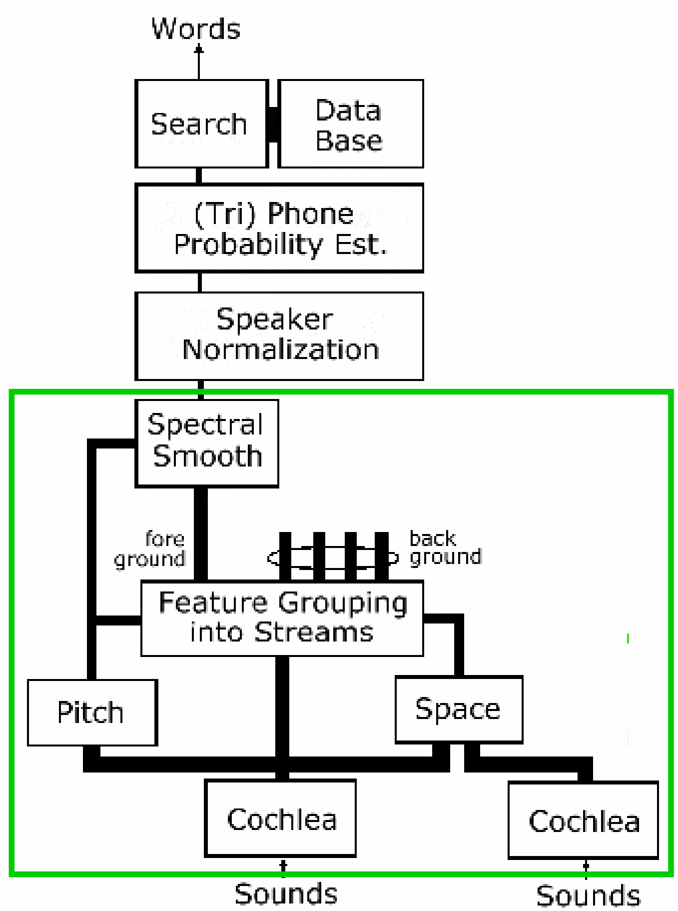


- Cortical functions: extensive pattern match, hypothesis generation and pruning, object tracking, HMM/Viterbi search, associative memory
- High-res feature detection, cross- and auto-correlation, and post-processing
- High-resolution sensory pre-processing

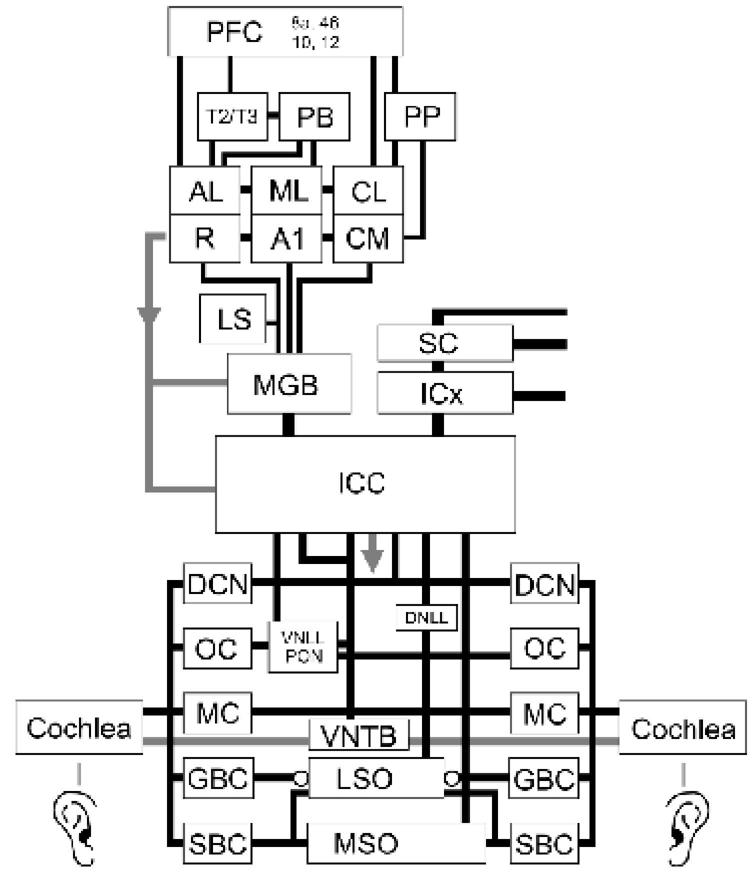
Commercial Example:
Stream Separation for Speech Recognition and Telecom



Single-source Recognition



Multi-source Separation and Recognition



AUDIENCE

Real-Time Demos

AUDIENCE™

Audience Introduction

- Fabless semiconductor and embedded software company with offices in Mountain View, CA
- Series B closed in April ‘06, $15M – total $25M to date
- First to market with commercial grade noise suppression based on the human hearing system
- Provides dramatic & reliable suppression of highly non-stationary (fast changing) noise sources such as another voice, music or ringtones
  - 20dB to 25dB noise suppression (non-stationary & stationary noise)
- Prominent investors & board members
  - New Enterprise Associates
  - Tallwood Venture Capital
  - Vulcan Capital
  - Carver Mead

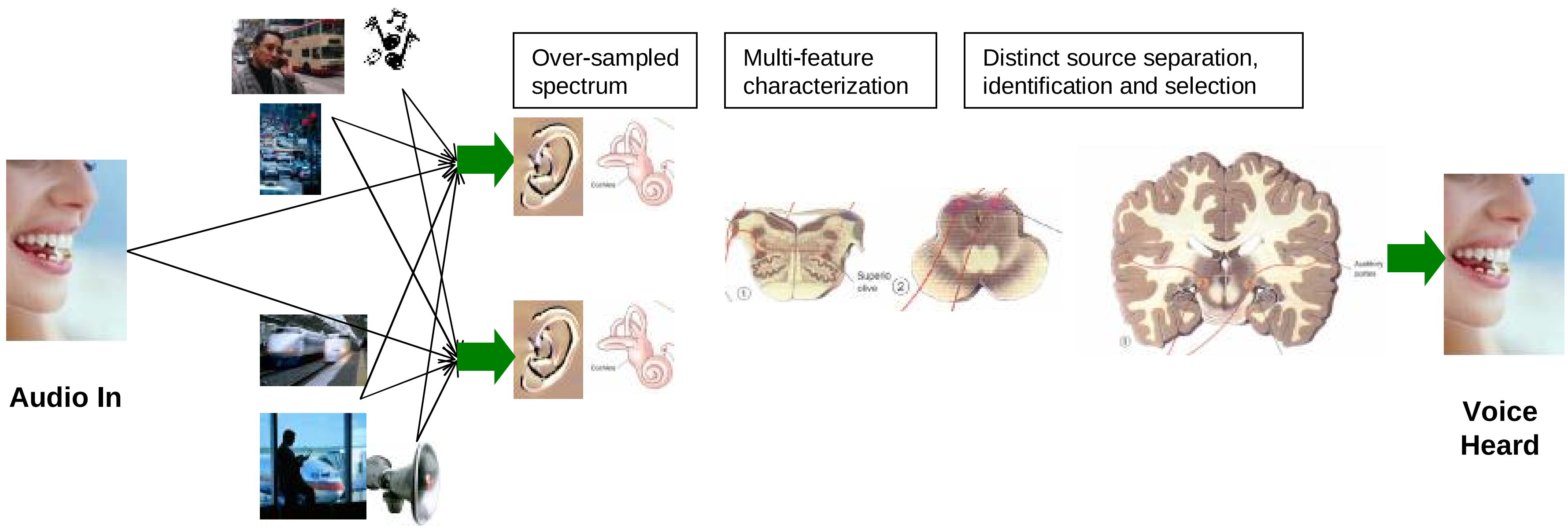
AUDIENCE™

Audience Background

- Management
  - Peter Santos – President & CEO, Director
    - LSI Logic, Voyan, Barcelona
  - Lloyd Watts, PhD. - Founder, Chairman & CTO
    - Interval Research, Synaptics, Arithmos
  - Bill Hoppin – VP Sales and Business Development
    - Celeritek, Accelerant Networks, Synopsys
- Outside Directors/Observers
  - Forest Baskett, New Enterprise Associates — Director
  - George Pavlov, Tallwood Venture Capital — Director
  - Steve Hall, Vulcan Capital — Director
  - Carver Mead — Director
  - Patrick Chang, TSMC/VentureTech — Observer
- Advisors
  - Dr. Lawrence Rabiner (Rutgers, ex-AT&T Bell Labs)
  - Dr. Robert Colwell (ex-Intel Pentium architect)
  - Dr. Vladimir Cuperman (UCSB)
  - Dr. Hynek Hermansky (OGI)
  - Dr. Abeer Alwan (UCLA)
  - Ray Kurzweil

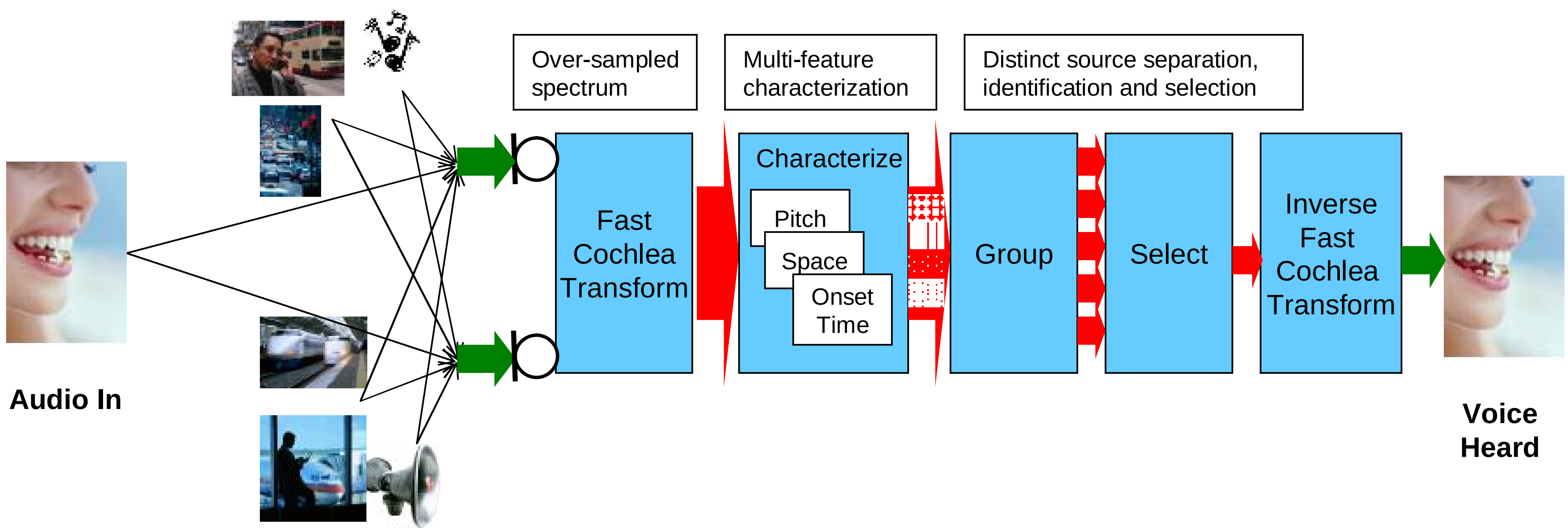
AUDIENCE™

The Biological System



The Audience System

Near Complete Suppression of Difficult, Distracting Non-stationary Noise
(as well as Stationary Noise)



Slowly Changing *Stationary* Noise

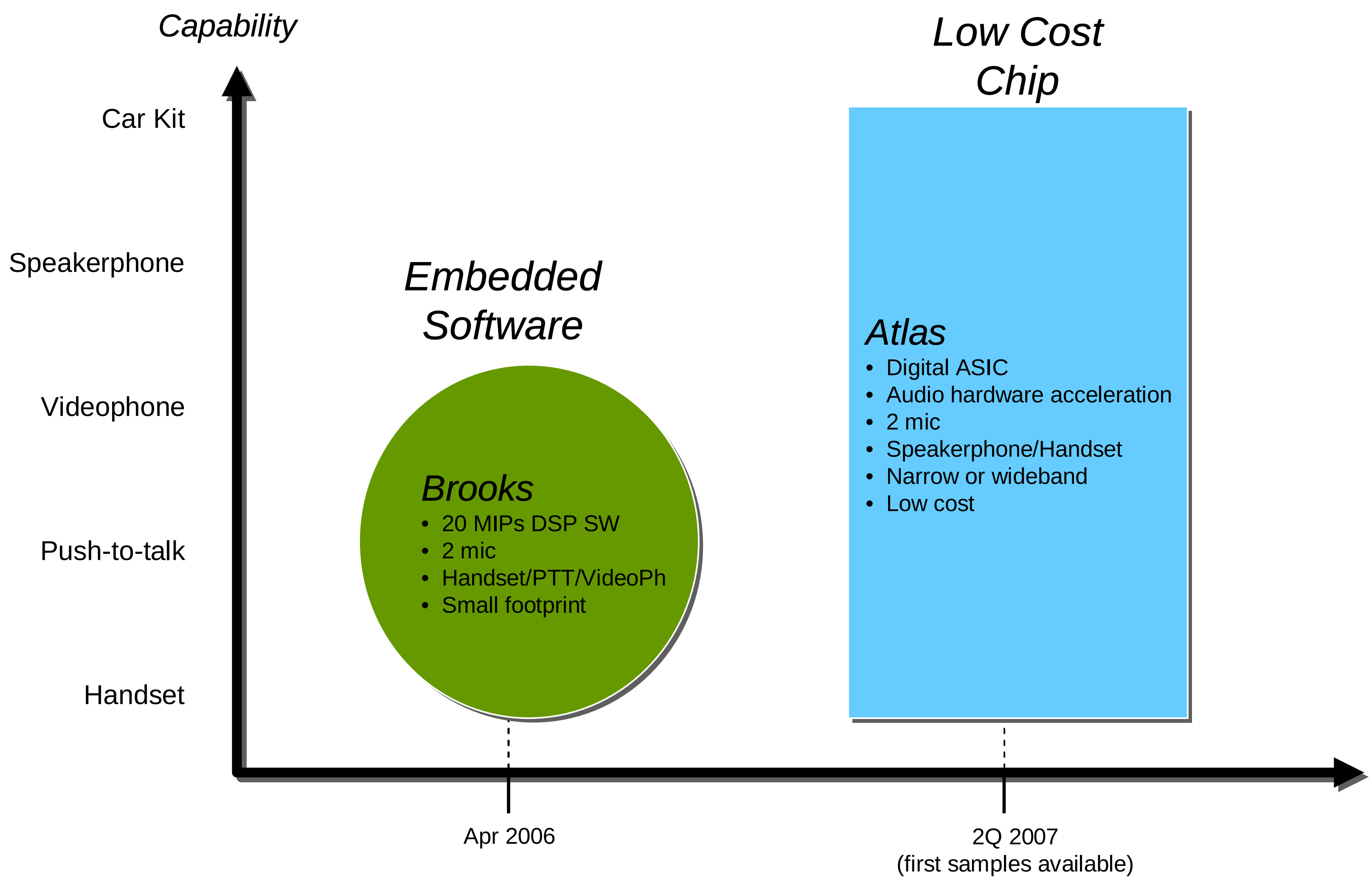
Fans, Crowds, Wind

Fast Changing *Non-Stationary* Noise

Sirens, voices, music
PA systems, trains

AUDIENCE™

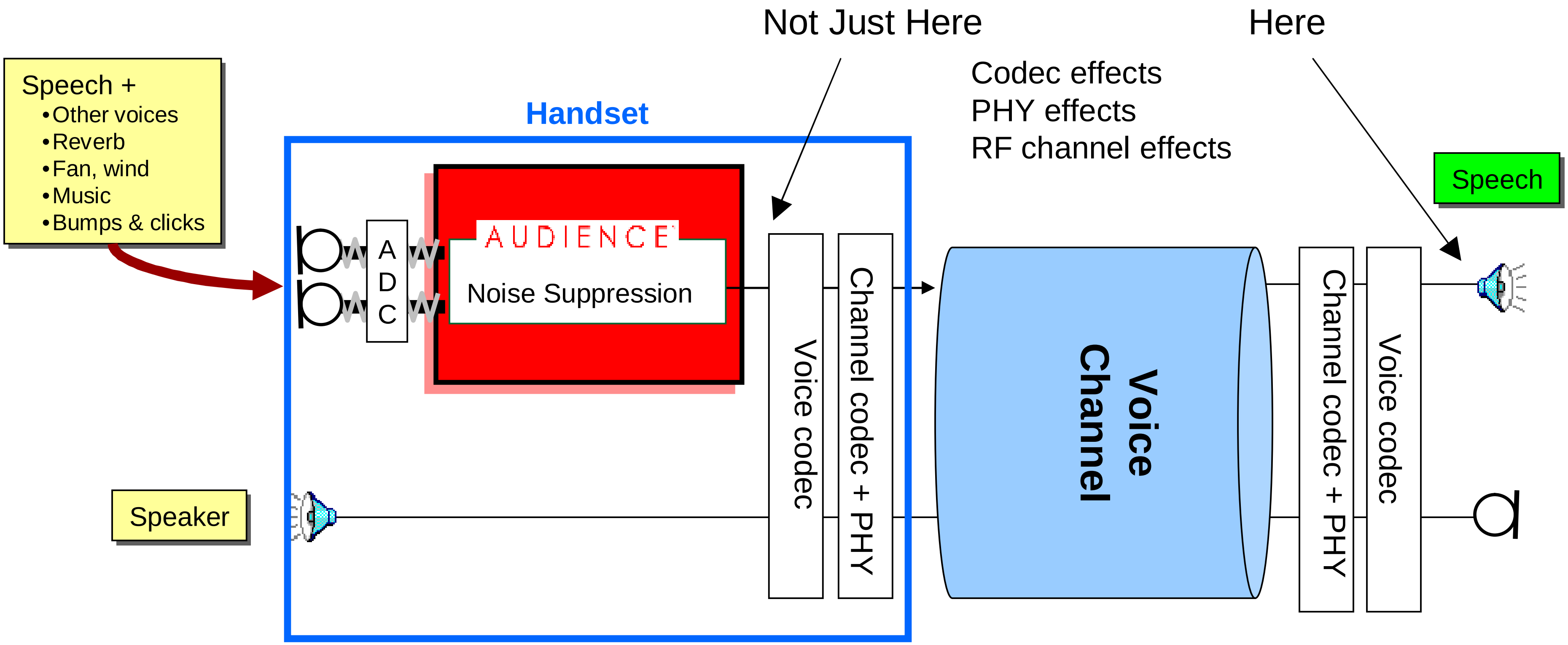
Audience Product Roadmap



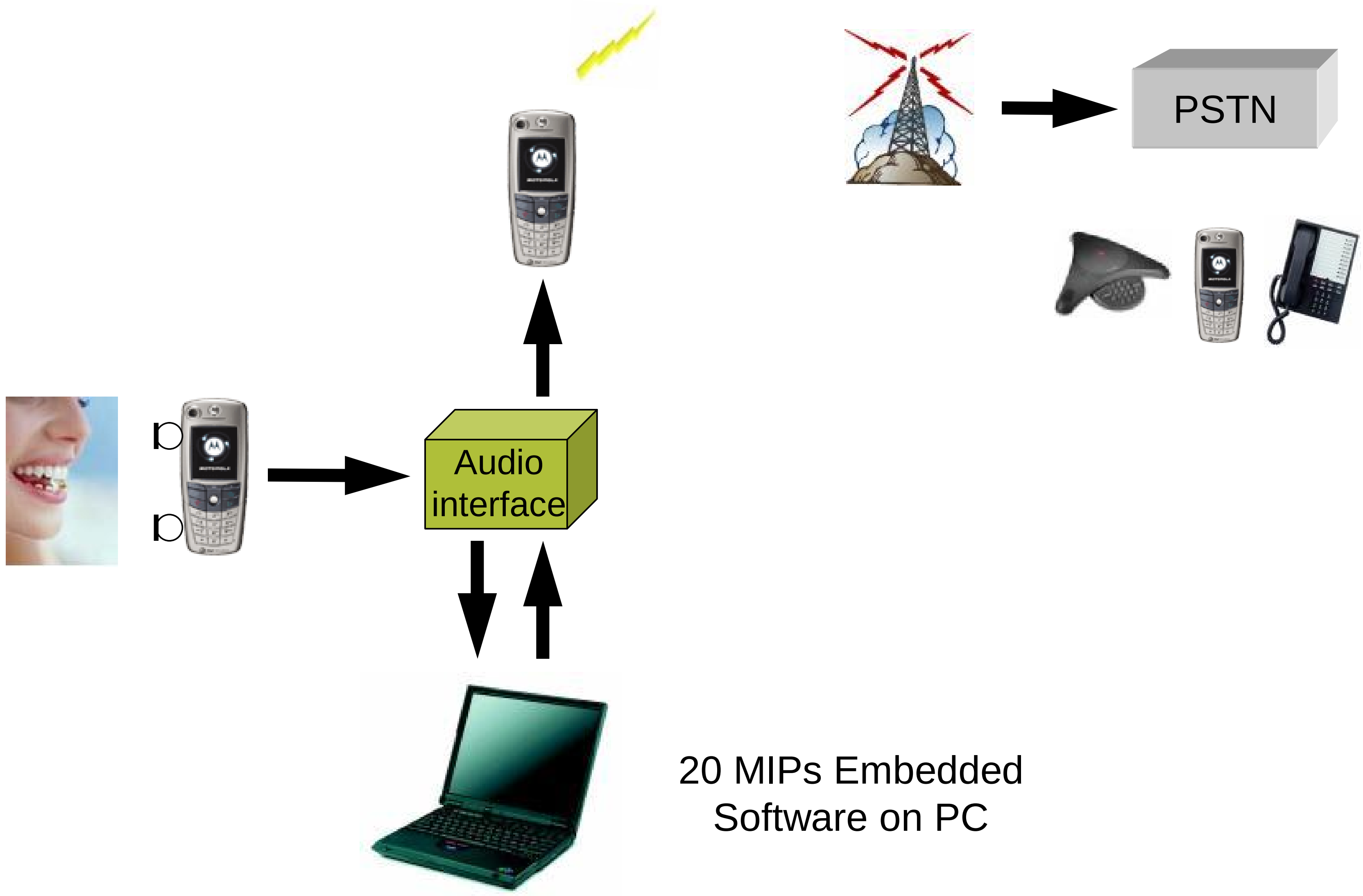
Audience Voice Quality Enhancement in Mobile Terminal

Audience system level optimization

1.) Aligned to specific terminal acoustics
2.) Optimized for performance where it matters: in the system



Audience real-time demonstration setup

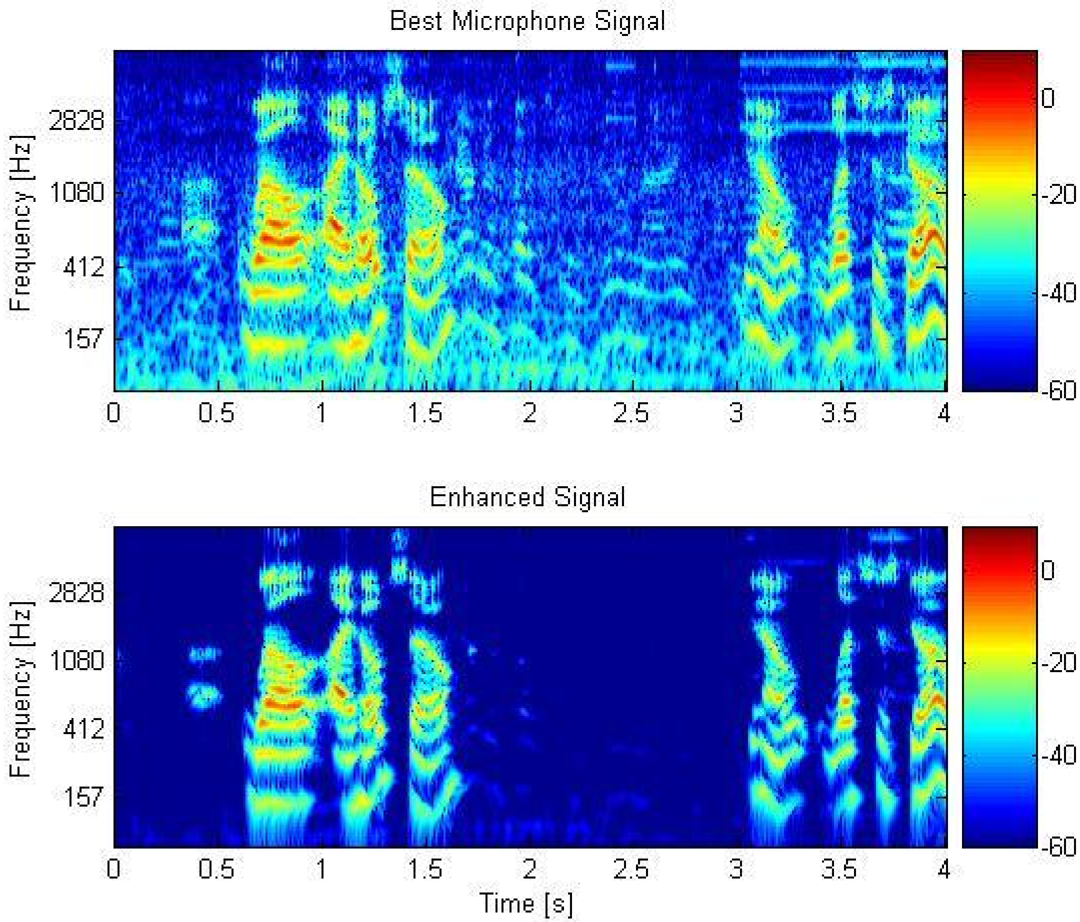


AUDIENCE

Audience real-time demonstration setup

Handset demonstration Streetnoise

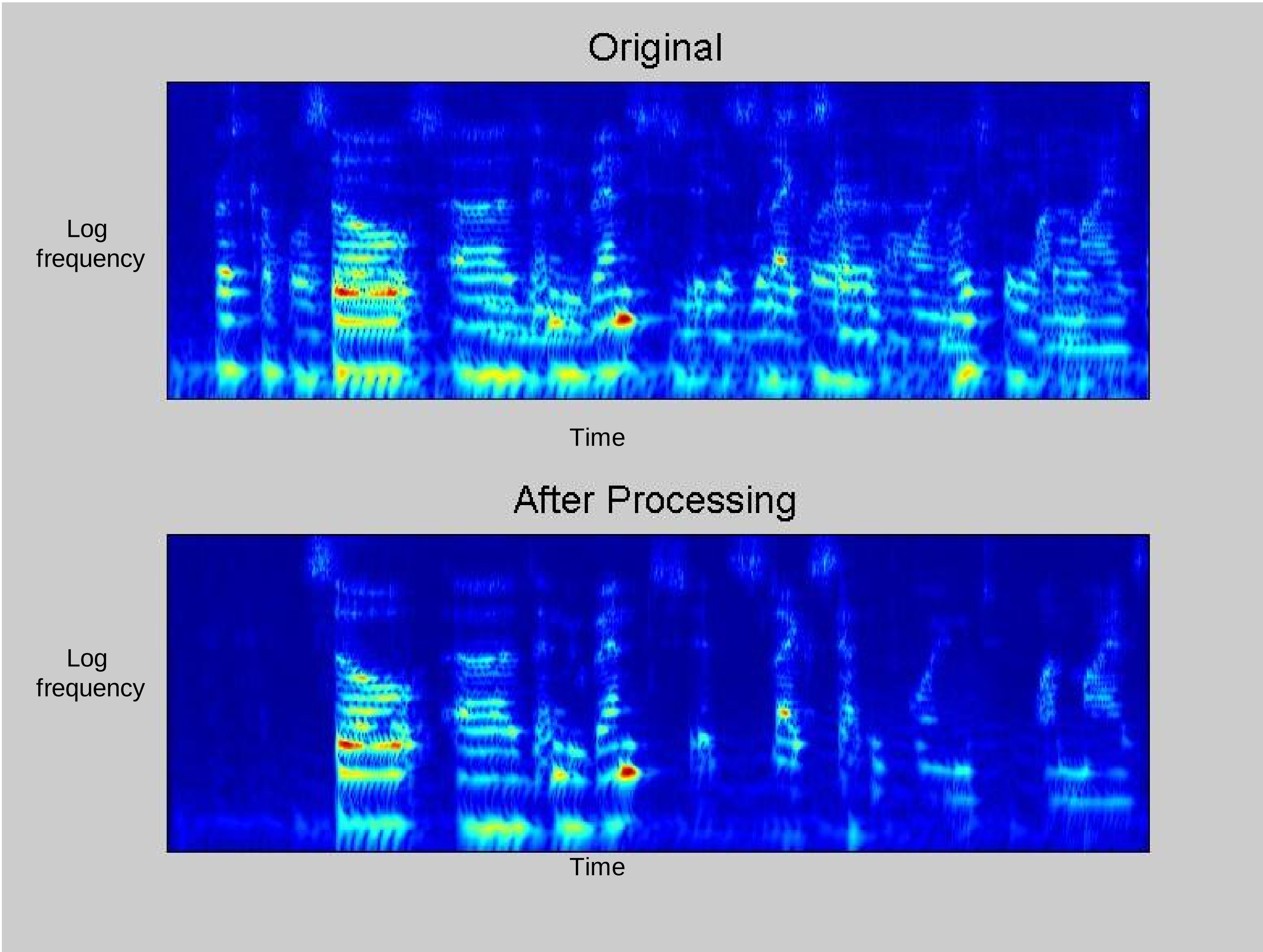
Brooks
Near Mic
Embedded SW



AUDIENCE

Speakerphone demonstration Competing voice

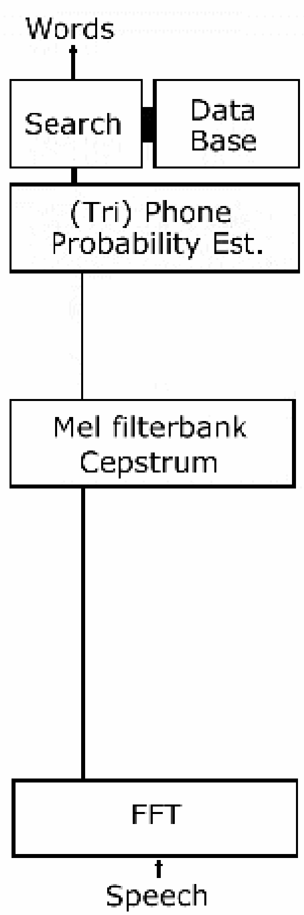
Atlas
Far Mic
ASIC



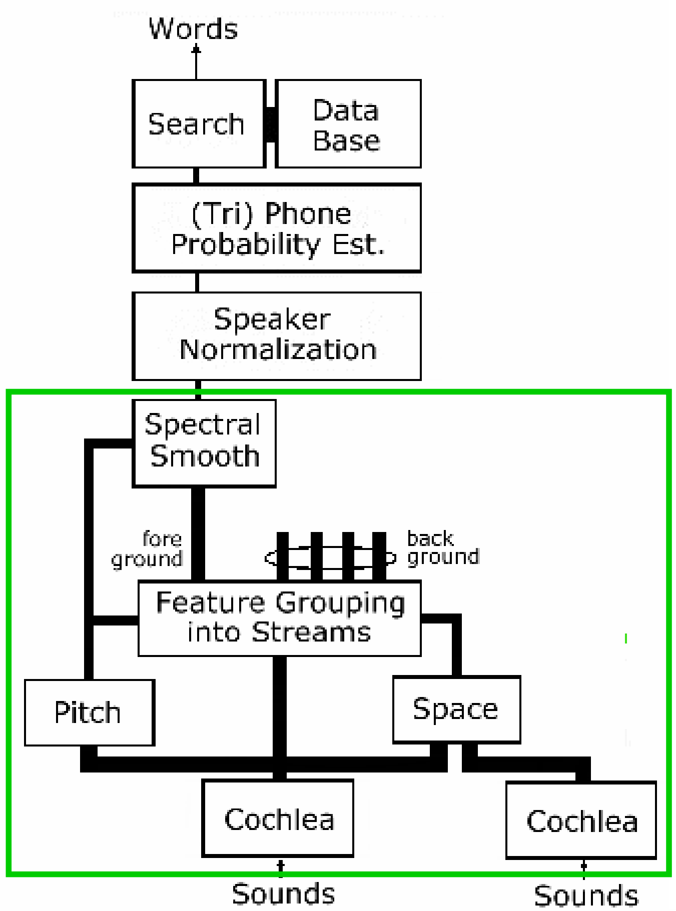
Recap and Conclusions

- Can we build a machine that hears the way human beings do?
  - Original passion: music transcription
  - Reverse-engineer the auditory pathway, based on neuroscience
  - Do we know enough about the brain? Are computers capable enough?
- If so, can we build a commercially successful company out of it?
  - Can we raise the money (i.e., convince the investors)?
  - What application to shoot for?
    - Music Transcription? No…
    - Speech recognition? No…
    - Noise suppression for Cell-phones? Yes!
  - Building a team, really executing
  - Is it a chip company, or a software company?
  - Investors, Customers, Employees, Advisors all have to see short-term progress, and long-term return

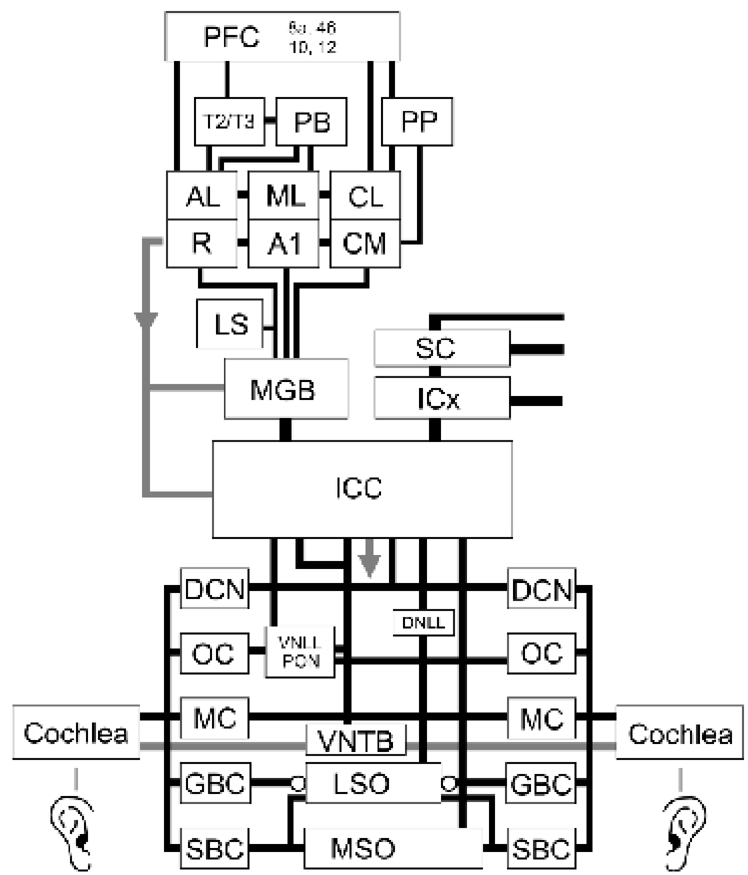
Commercial Example:
Noise Robustness for Speech Recognition and Telecom



Single-source
Recognition

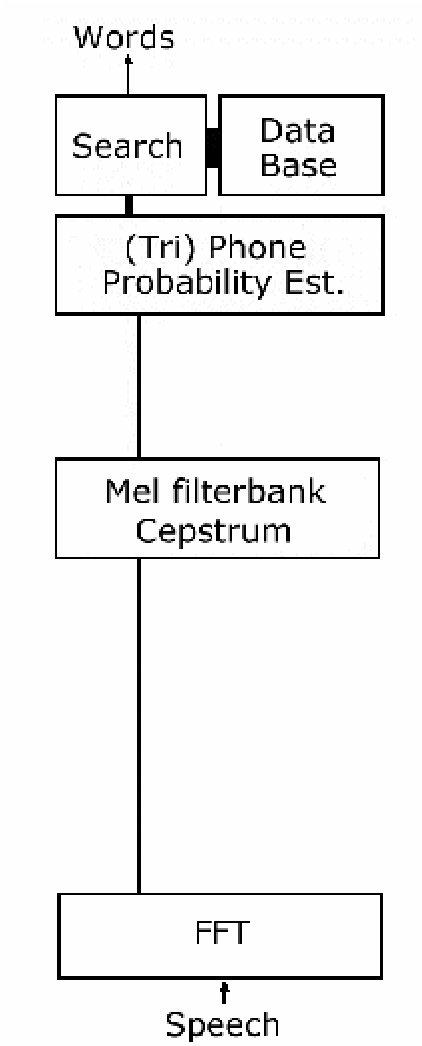


Multi-source
Separation and Recognition



AUDIENCE

Resolution Requirements are different for Multi-Source Separation than Single-Source Recognition



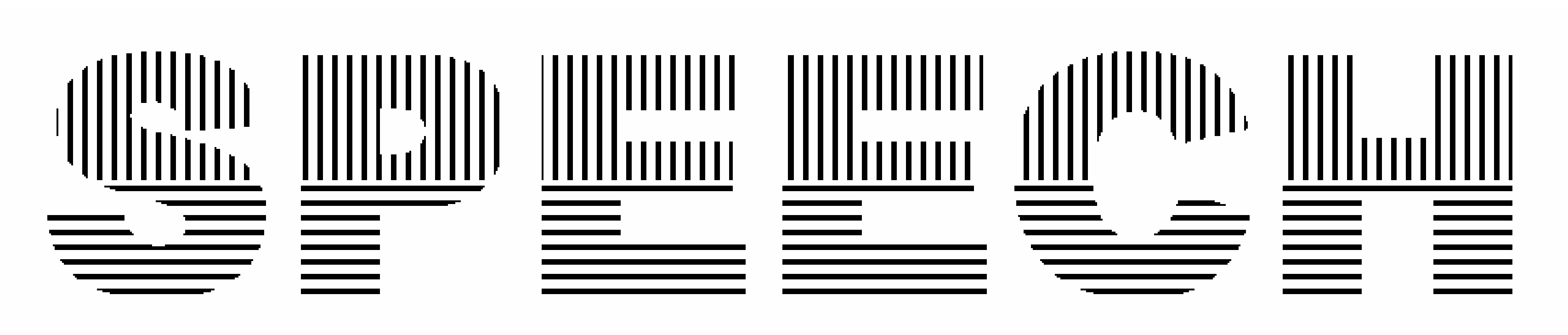
"The selection of the best parametric representation of acoustic data is an important task in the design of any speech recognition system. The usual objectives in selecting a representation are to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data and to enhance those aspects of the signal that contribute significantly to the detection of phonetic differences… Compact storage of the information [is] an important practical consideration."

-- Paul Mermelstein (inventor of mel-frequency cepstral coefficient), 1980.

"The very purpose of phonetic classification in ASR is a significant reduction of the information carried by the speech signal. Thus, the front end processing should be supportive of this task."

-- Hynek Hermansky, 1998.

A Visual Analogy

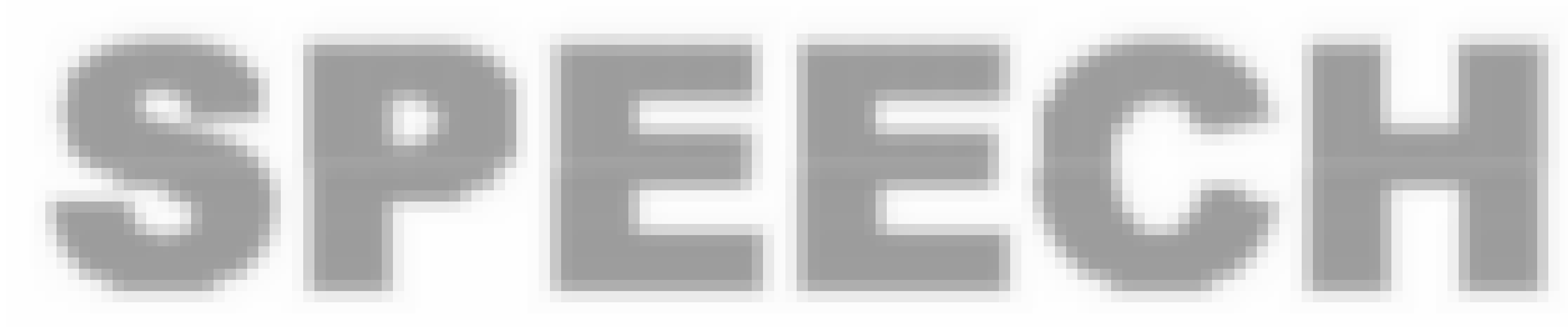


- Voiced speech contains fine structure due to pitch, speaker ID…

A Visual Analogy

- Voiced speech contains fine structure due to pitch, speaker ID…
- Blurring eliminates fine structure, making pattern matching easier…

A Visual Analogy



- Voiced speech contains fine structure due to pitch, speaker ID…
- Blurring eliminates fine structure, making pattern matching easier…
- And justifies downsampling, which reduces downstream system cost.

Here is a list of states:
Alaska, Ohio, Florida,
Nebraska, Delaware,
Kansas, Wisconsin.

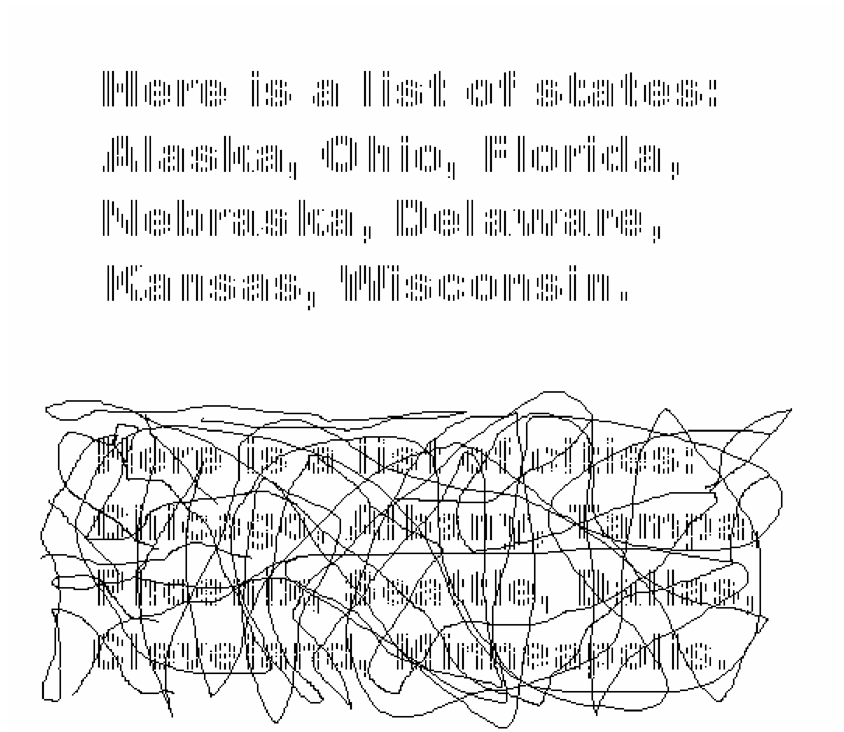
Here is a list of states:
Alaska, Ohio, Florida,
Nebraska, Delaware,
Kansas, Wisconsin.

Here is a list of cities:
[illegible]
[illegible]
[illegible]

A Visual Analogy

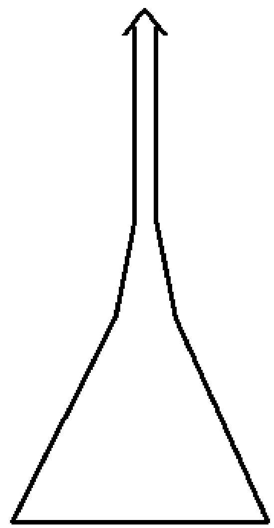
Low-Res OK for single-source

Hi-Res Needed for multi-source



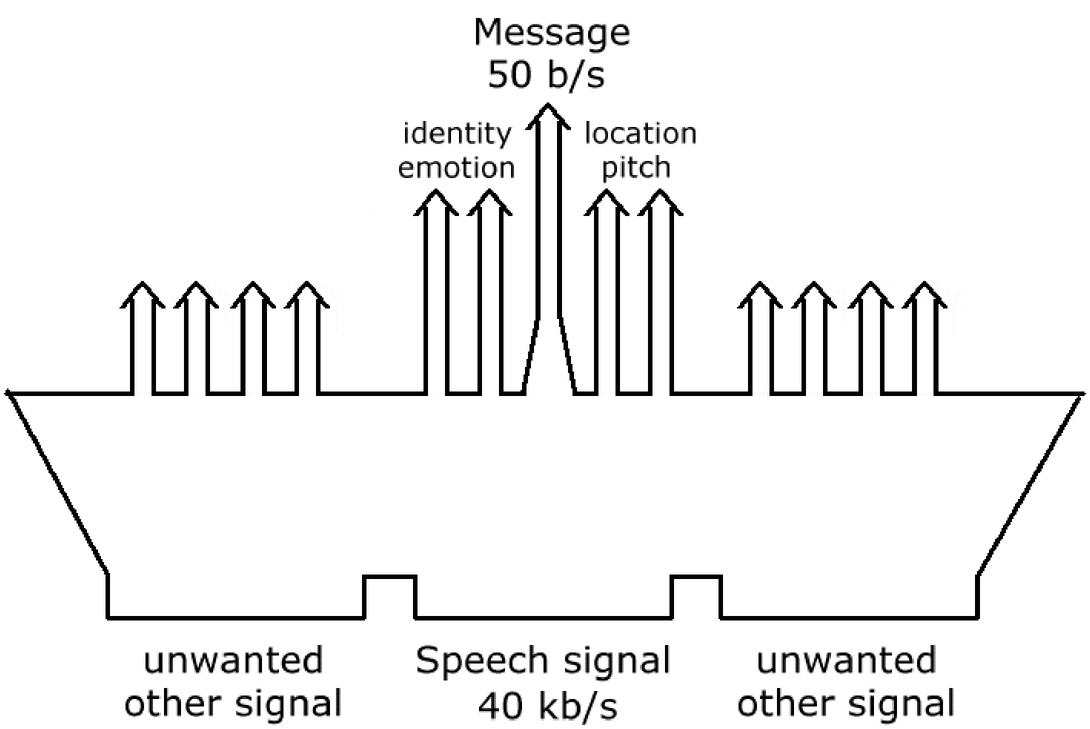
AUDIENCE™

Information Flow for Single-Source



(Hermansky, 1998)

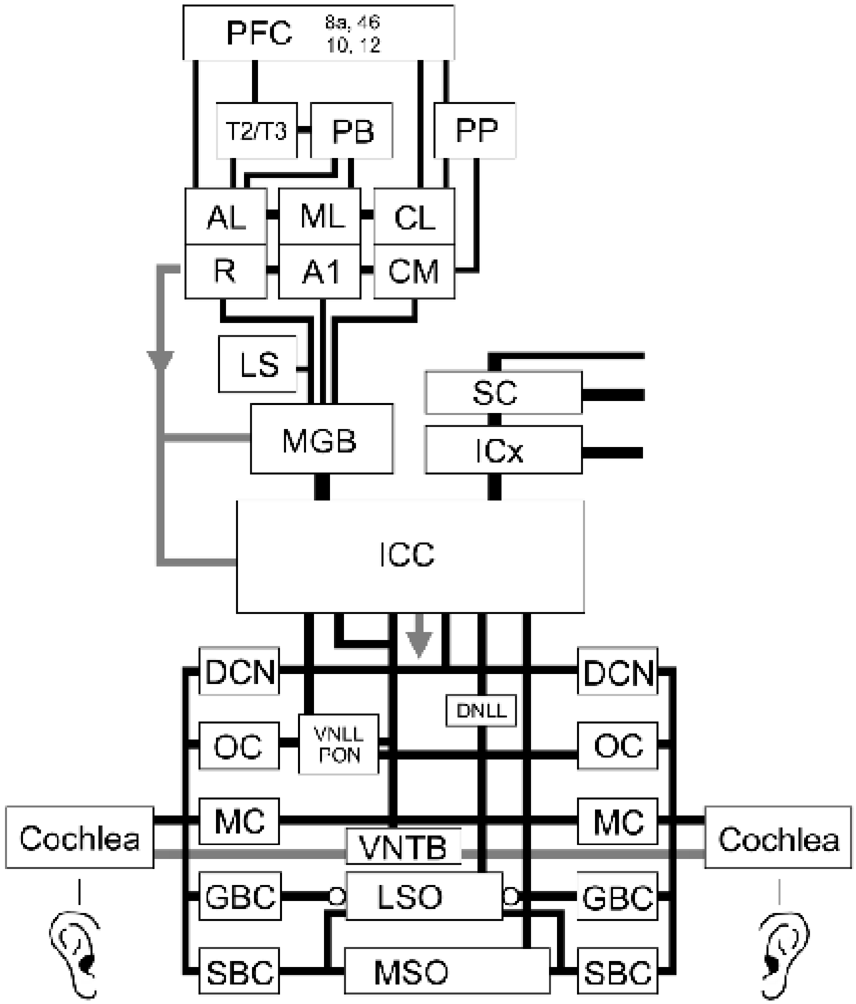
Information Flow
for Multi-Source



Pruning

Hypothesis generation

Architectural Implications



- Cortical functions: extensive pattern match, hypothesis generation and pruning, object tracking, HMM/Viterbi search, associative memory
- High-res feature detection, cross- and auto-correlation, and post-processing
- High-resolution sensory pre-processing

AUDIENCE™

Architectural Implications

Human-Level Performance on Real-World Multi-Input Sensory Processing will require:

- Further algorithm development to define robust computational pipeline (analogous to graphics rendering pipeline)
- Real-time Hearing will likely require 10-100 GOps range
- Will need efficiencies better than 3 Gops/W for commercial acceptance in many applications
- Priority on fast-turnaround design, high-res visualization, real-time operation, validation on large data sets for robustness, low latency.
- Power consumption, cost reduction are later-stage optimizations once the key operational principles are understood
- Will favor parallel, pipelined, low-clock-rate, low-voltage hardware-accelerated architecture to bring power consumption to reasonable level, but power consumption will trade against chip area and flexibility
- Best implementation model may be GPU – parallel compute for dedicated processing pipeline