SpeechWorks® solutions from ScanSoft®

Productivity without boundaries

# Embedded Speech Recognition: State-of-art & Current Challenges

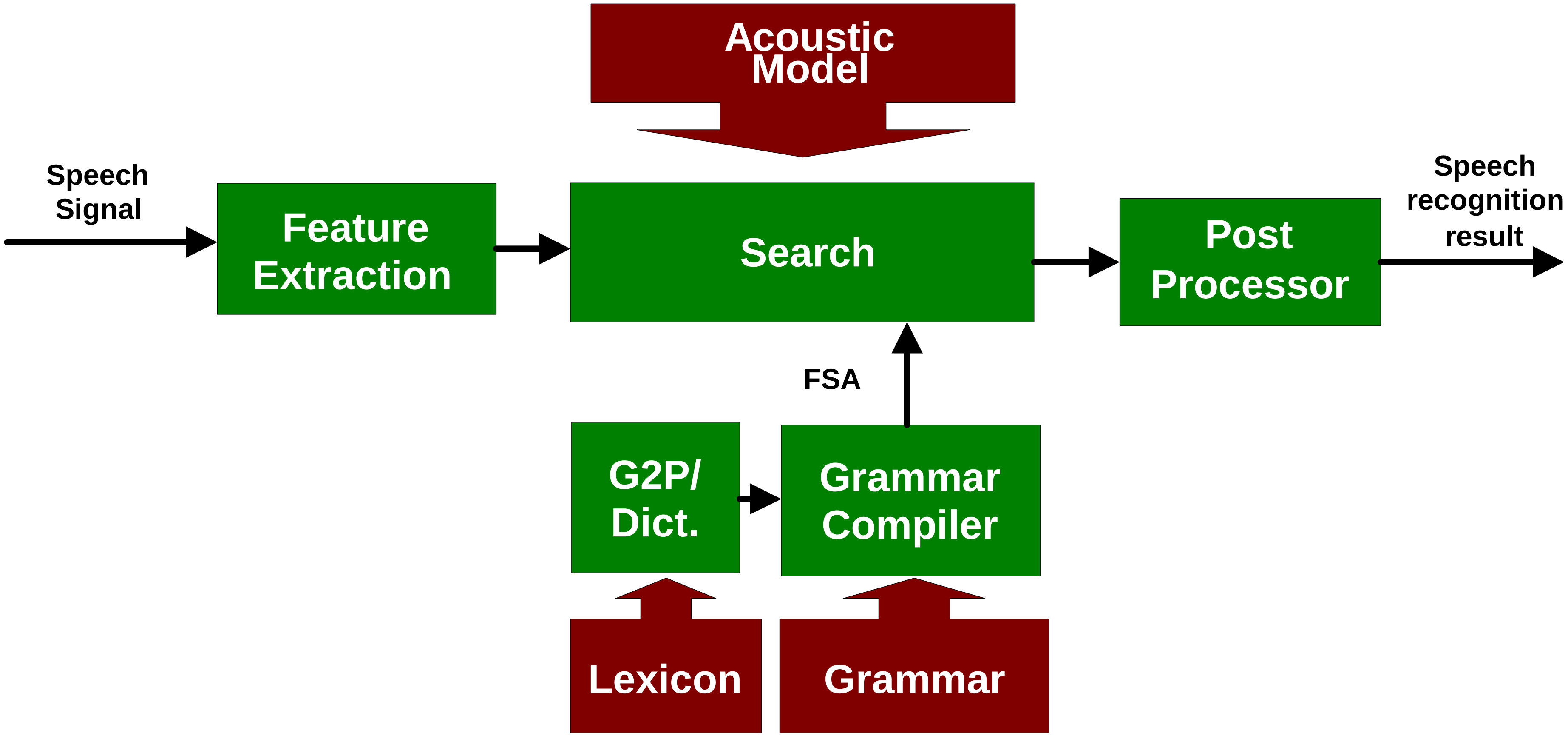
*Presented by:*

Dr. Ir. Christophe Couvreur

# Plan

- A Refresher on ASR Technology
- Embedded ASR: Platforms & Applications
  - Automotive
  - Mobile
  - Game
- Constraints & Influence on State-of-the-Art
- Where is the Multi-Modality Today?
- Demos
- State-of-the-Art vs. Cutting Edge
- Research Challenges
- Conclusion & Discussion

# A Typical Speech Recognizer

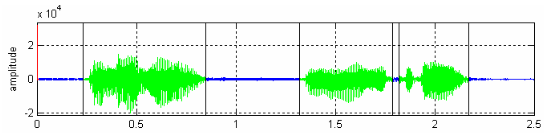


# Feature Extraction Technology

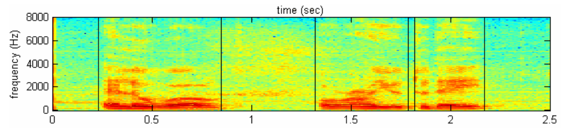
- Extract features characteristics of the speech signal:
  - Time-frequency analysis of the input signal
- Remove unwanted influences
  - Noise cancellation
  - Normalization (e.g. for differences in microphone frequency responses)
  - Pitch
    - → keep only evolution of spectral shape (speech formants)
    - Asian language: extract pitch/tone information!
- Most common Mel-Frequency Cepstral Coefficients (MFCC)
  - Alternative: RASTA-PLP
  - Work from speech codec parameters on cellphones

# MFCC Front-End

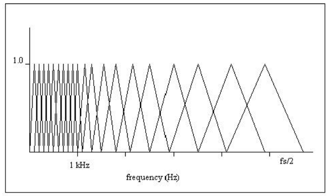
PCM Input



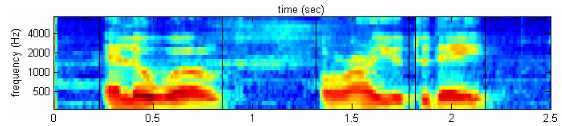

**Time-frequency analysis (Fourier Transform)**




**[Noise cancellation] Perceptual band-pass integration & rectification**




**Cepstral transformation Normalization**




**Frame Stacking (δ,δδ)**

**Linear Transformation Dimensionality Reduction**

Feature vectors stream

# Acoustic Modeling Technology

SpeechWorks® solutions from ScanSoft®

- Statistical model of speech pronunciation (sequence of phonemes)
- Standard acoustic modeling:
  - HMMs (hidden Markov models)
    - Represent frequency and temporal statistical variations of specific sounds
  - Mixture of Gaussian pdfs is most common today
    - Parametrization: mean, covariances, weights
    - Many variants ("tying")
  - Context-dependent modeling w/ generalized tri-phones
    - Phonetic context decision tree to avoid combinatorial explosion of # of context-dependent phones
- Alternative: Hybrids Multi-layer Perceptrons & HMMs

# Mixture of Gaussian Model



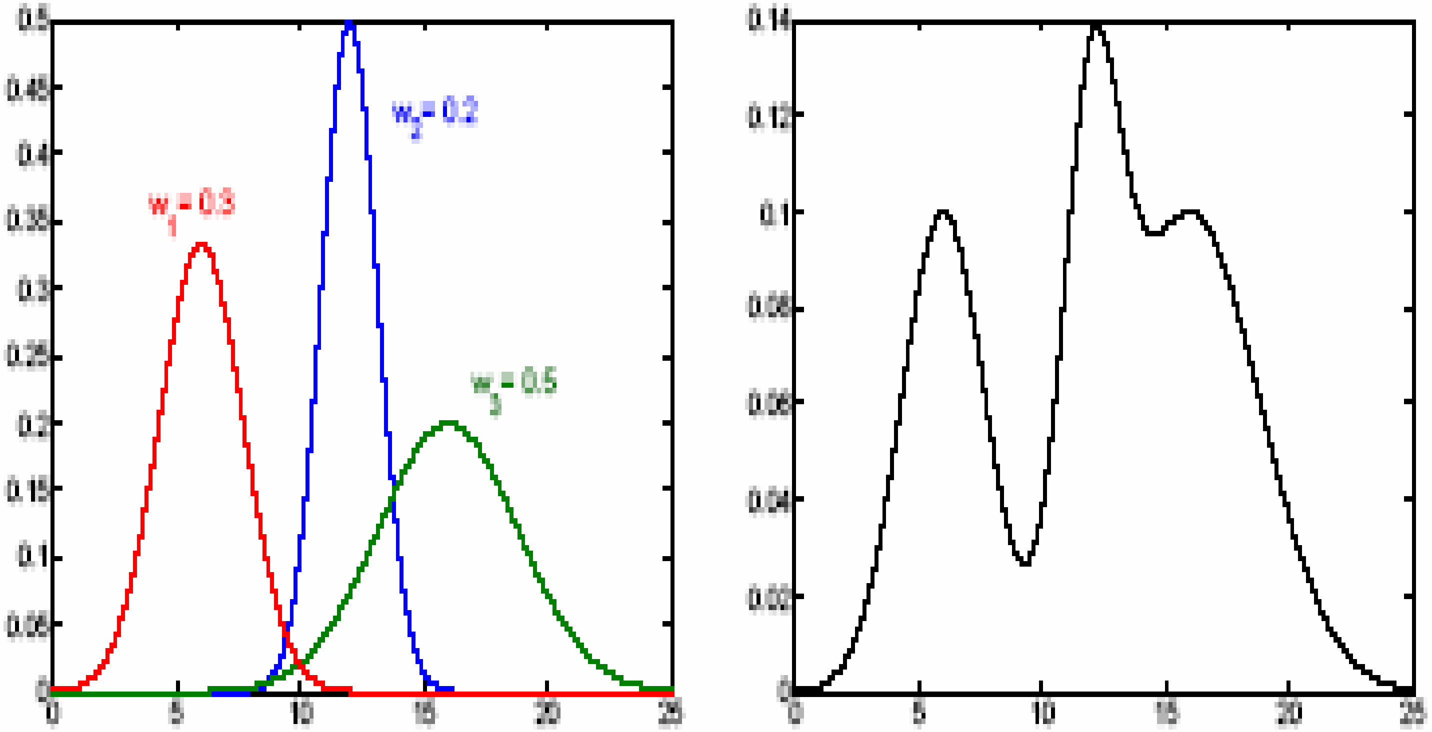
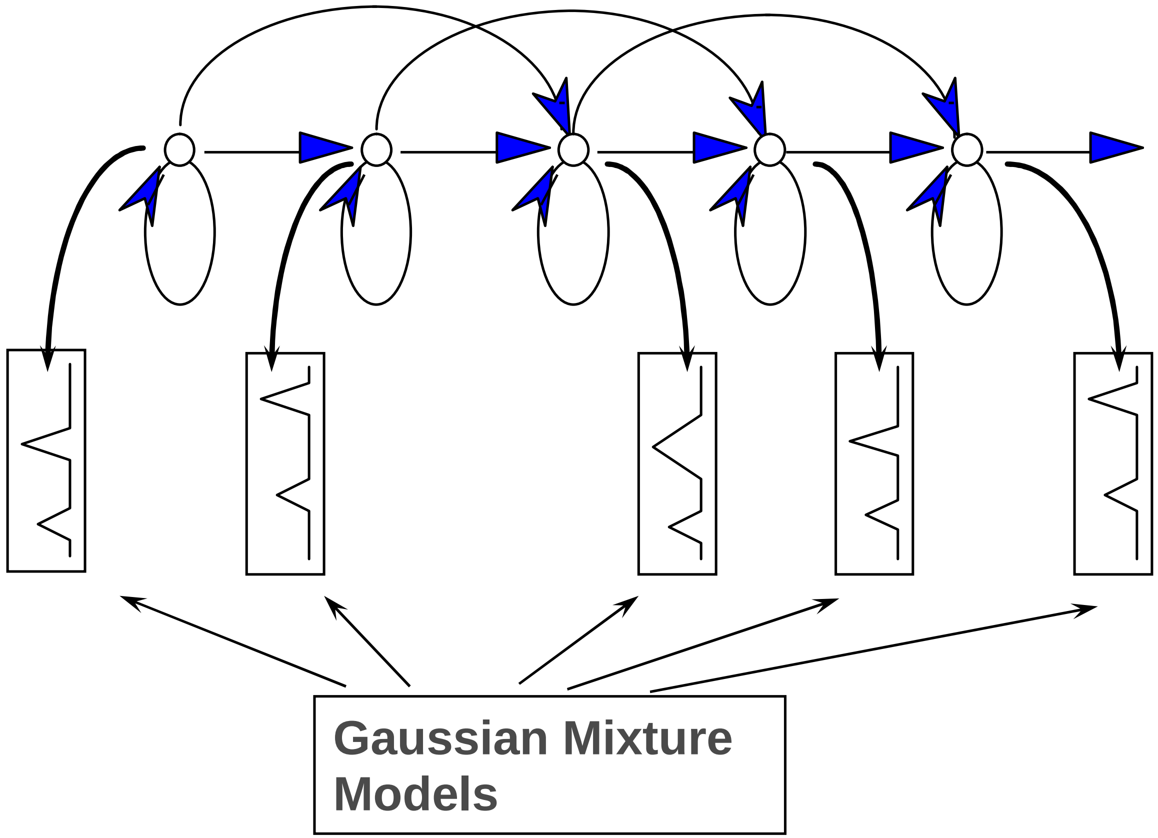



Figure 1: One dimensional Gaussian mixture pdf, consisting of 3 single Gaussians

# Hidden Markov Model

Model for 1 phone in context



# Phonetic Context Decision Tree

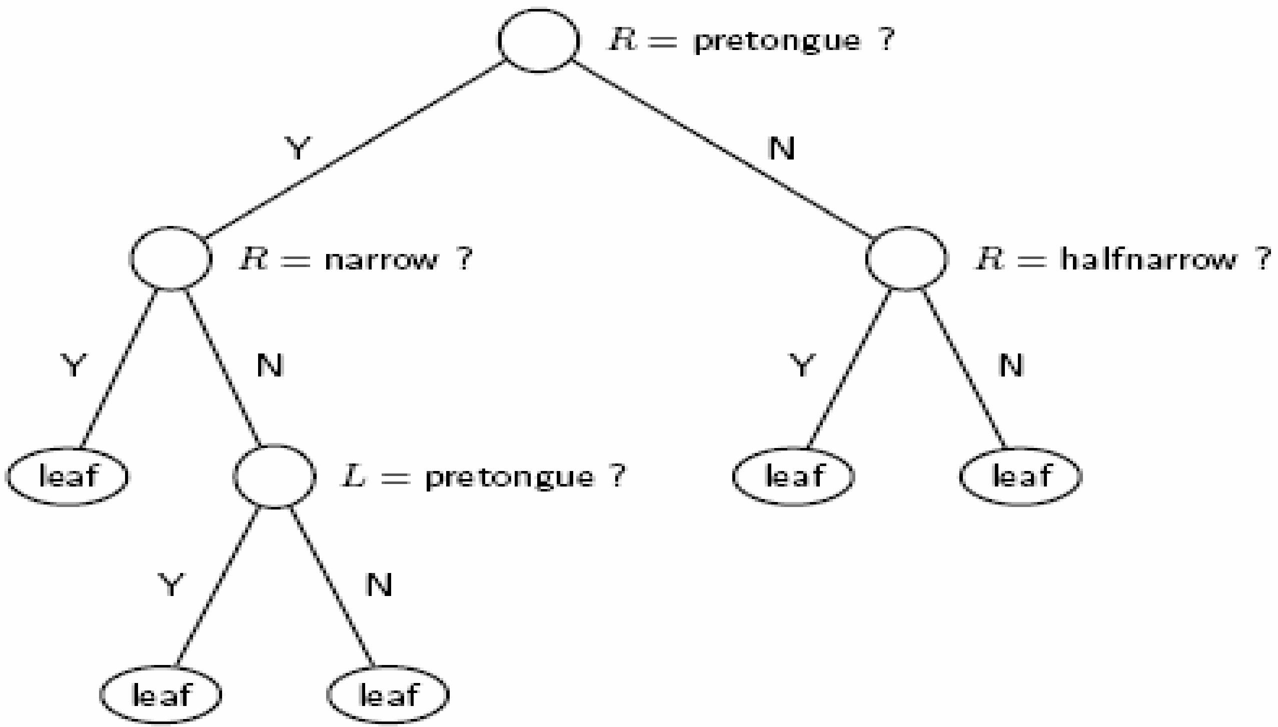



Figure 2: A phonetic decision tree for the second state of /k/.

# Acoustic Modeling Technology (cont'd)

SpeechWorks® solutions from ScanSoft®

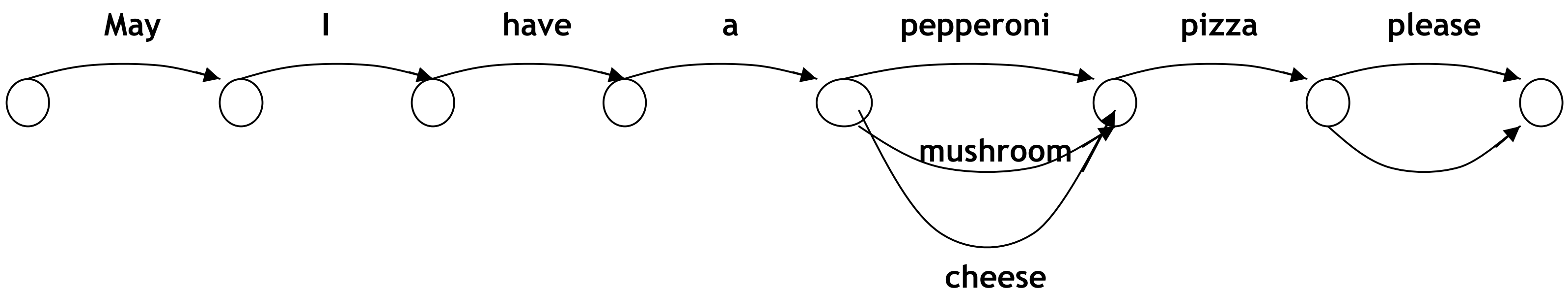
- **Speaker independent** vs speaker-dependent
- Trained on (very) large corpora of speech:
  - "There's no data like more data"
  - 100's or 1000's of speakers per languages
  - Adequate phonetic coverage
  - Environment & domain-matched if possible (field-collected)
- Speaker adaptation:
  - Adjust model parameters to specific speaker
  - MLLR (Maximum-Likelihood Linear Regression)
  - MAP (Maximum A Posteriori estimation)
- Rejection of out-of-vocabulary speech/noises.

# Lexicon & G2P Technology

- Text → phoneme conversion:
  - Example: `Couvreur` → `#ku.'vRE+R#`
  - Lexicon (manually transcribed dictionaries)
  - Grapheme-to-Phoneme conversion algorithm
  - Shared w/ TTS systems
- G2P:
  - Rule-based:
    - Manually crafted or trained from corpora
  - Statistical systems
    - Decision trees
    - HMM-based
  - Combination of the two approaches, combined w/ "backbone" lexicon.

# Grammar & Language Model Technology

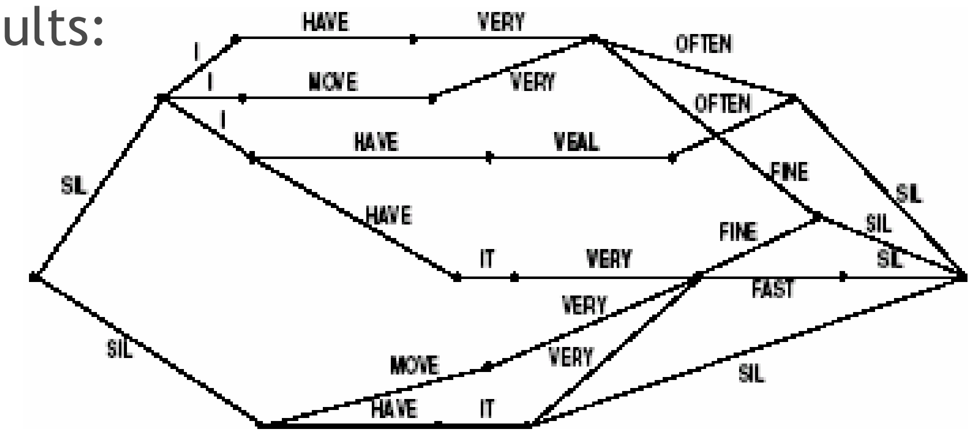
- Formalism to describe acceptable sequence of words
- Grammar:
  - Explicitly describe all valid transitions
    `<start> :  may I have a (mushroom | pepperoni | cheese) pizza please ;`




  - Manually crafted, with dynamic slots
- Statistical language model (SLM):
  - All transitions are valid, some more likely than others.
  - Tri-grams (N-grams): p $(w_1 \mid w_2\ w_3)$
  - Example: p('rose'|'stock prices') vs. p('fell'|'stock prices')
  - Trained on relevant text data

# Search Technology

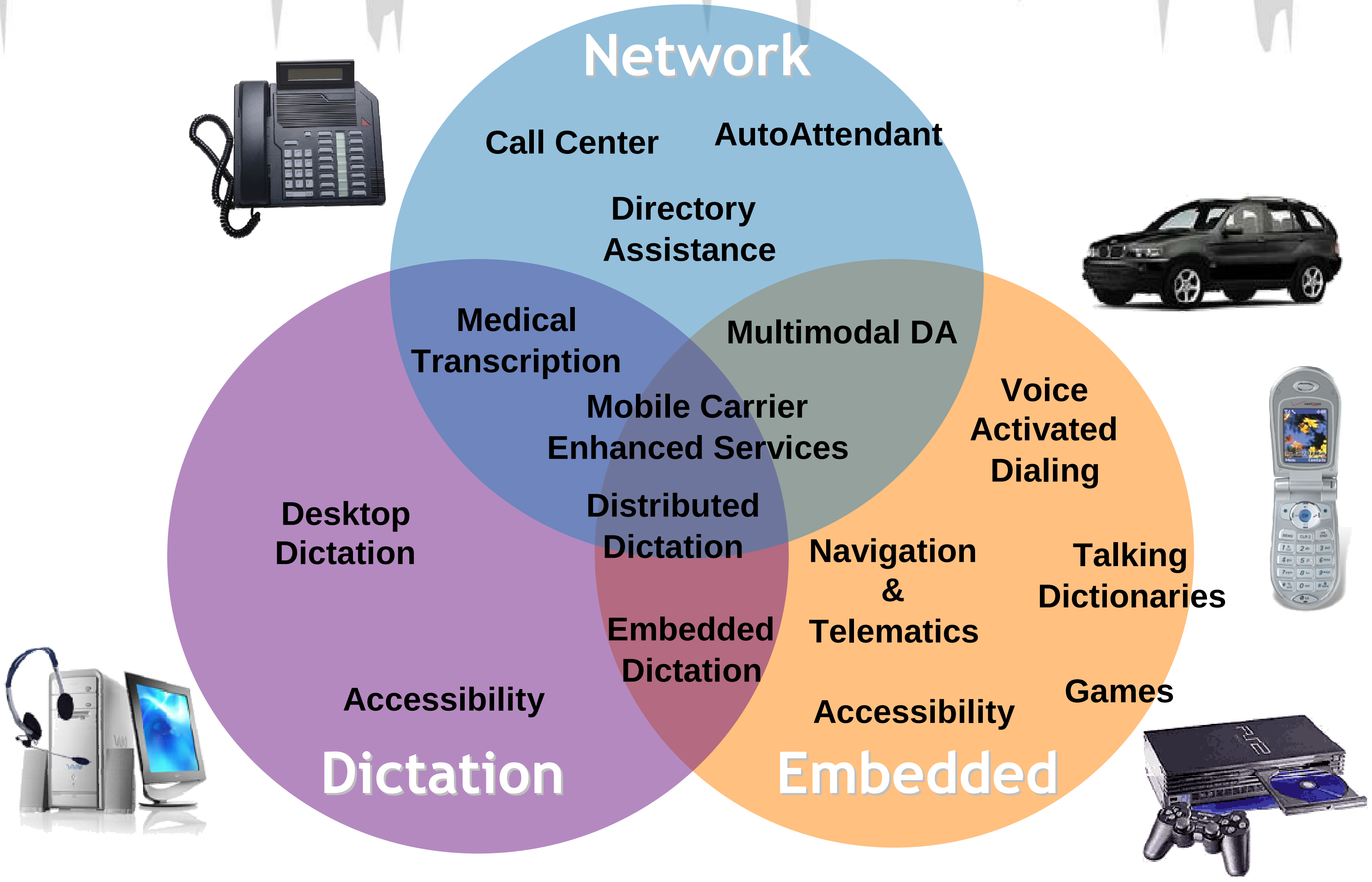
- Typically domain specific & optimized.
- Essentially variations on the Viterbi dynamic programming (DP) algorithm
  - Except for IBM's A*-search (envelope search) technique
- Optimizations to minimize CPU and memory impact
  - Pruning technique (beam search)
  - Pre-filtering or multi-pass techniques: using approximation in first pass then refine in subsequent passes on limite set of hypotheses
- Can supply alternative results:
  - N-best lists
  - Lattices



# Post-processors

- Applications-specific, often tied into the grammar/search approach.
- Examples for SLM-based systems:
  - Inverse Text Normalization in dictation
    - "twenty first of july two thousand five" → 'July 21, 2005'
  - Robust parsing for information retrieval systems
    - "Can you tell me when the **United Airline** plane for **Paris** leave **tomorrow morning**, please?"
- Examples for Grammar-based systems:
  - Spelling for destination entry in GPS system
    - "S T U T G A R D" → 'Stuttgart'
  - Normalization of output format (NLU)
    - "Call 9 1 1" or "Call 9 eleven" → 'Call 911'
  - N-best reordering based on a priori knowledge (e.g. usage history)
- Dialog or fusion layers on top of the recognizer can be viewed as additional post-processing steps

# Speech Markets



# Embedded ASR Markets

SpeechWorks® solutions from ScanSoft®

- Automotive
  - Navigation systems
  - Telematics units
  - Hands-free car-kits
- Handhelds
  - Handsets
  - SmartPhones
  - PDA
- Game consoles
- *Other Devices*
  - *Military*
  - *Industrial (e.g. wharehousing)*
  - *Language Learning*

# Today's Speech Applications?

SpeechWorks® solutions from ScanSoft®

- Moore's law makes it possible to deploy more complex speech technologies on embedded platforms
- Driving factors:
  - Technological performance of platforms
  - Customer acceptance of speech
  - Price: bill-of-material (BOM)
  - Length of development cycles/time-to-market
    - 2-5 years in automotive
    - shorter in handhelds (~1 year)

# Volume Applications – Today

- Automotive:
  - Hands-free car kits, with continuous digit dialing & SD dialing
  - On-board C&C, including navigation system & telematics
- Handhelds:
  - SD dialing
  - SI name dialing & digit dialing
- Video Games
  - Simple C&C in games

# Volume Applications – Tomorrow

SpeechWorks® solutions from ScanSoft®

- Automotive:
  - BlueTooth car kits, with SI name dialing
  - Navigation with voice destination entry
  - Control of MP3 player (e.g. iPod)
- Handhelds
  - Broader availability of voice-activated dialing
  - Control of MP3 player
  - (Continuous) dictation of SMS & e-mails
  - Convergence of local & distributed speech processing
- Video Games
  - New interactive use of speech in games

# Platform Capabilities – Car-kit

- DSP Processor:
  - TI C54xx, TI C55xx, TI OMAP, Infineon TriCore, ADI Blackfin, …
  - 50—200 MIPS
- RAM:
  - Paginated internal memory: 64 KW
  - < 256—512 KB total
- No OS or very limited
- Fully allocated to speech (AEC/ASR)

# Platform Capabilities – Navigation

- RISC Processor
  - Example: Renesas SH4, Motorola MPC 5xxx, TI OMAP, etc.
  - 70-80% available for speech
- RAM:
  - 32—64 MB
  - 4—16 MB available for speech
- DVD storage
- Embedded OS
  - Windows CE or ~Unix (Linux, QNX, VxWorks)
- Often combined w/ Telematics unit (e.g. BMW i-Drive)

# Platform Capabilities – Handhelds

- Handset
  - Heavily segmented market
  - ARM7 or ARM9 processor: 30—200 MHz
  - RAM: 2—8 MB, with 256 KB—1 MB for speech (more for Dictation)
  - 8 → 16 kHz
  - Proprietary or Symbian OS
- SmartPhone / PocketPC
  - Intel Xscale, TI OMAP, ARM9 or ARM11 processor: 200-600 MHz
  - RAM: 16—64 MB, with up to 16 MB for speech (dictation)
  - Windows CE or Linux
- Flash storage

# Platform Capabilities – Consoles

- Game consoles:
  - Sony PS2 & PSP
  - Microsoft Xbox, (360)
  - Nintendo Game Cube
- RISC Processor
  - Plus floating-point coprocessor
  - 300-700 MHz, up to 4 Gflops
  - Speech < 5—10%
- RAM:
  - 24—64 MB
  - Speech < 500 KB—1 MB
- DVD storage
- Next Generation (Sony PS3, Microsoft Xbox 360)
  - PowerPC processors (Cell 7x 3.2 GHz for PS3)
  - 256 MB RAM

# Constraints Impact

*Platform constraints have an impact on technological choices*

- Examples:
  - DSP's with fast absolute distance computation → use of Laplacian densities
  - DSP's with paginated memory → use of 1.5-pass search/distance computation
  - Limited memory → off-line tools for grammar compilation/G2P + limited dynamic run-time modifications (FSM)
  - Limited resources → tuned acoustics models

# Constraints Impact (cont'd)

- Examples:
  - Limited CPU & memory → special DPs (digits loop, long item lists [multi-pass & « Fast-match », lexical tree], … )
  - Limited memory → pruning techniques
  - Real-time / vsynch processing → « hard » real-time processing (synchronous)

*Application-level constraints have an impact too*

- Examples:
  - VDE → pronunciation databases for toponyms
  - VAD → sharing of G2P with TTS

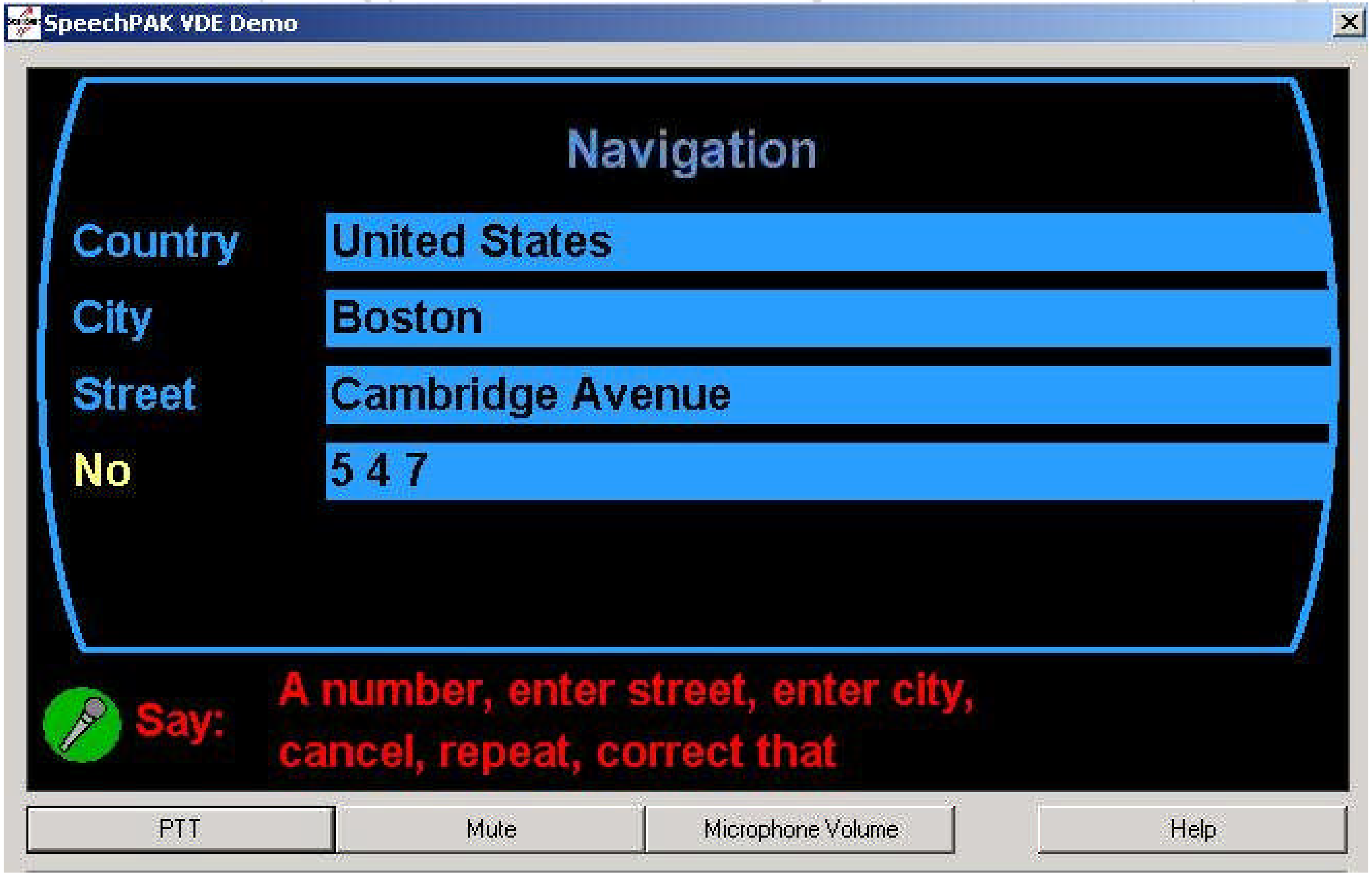
# UI Issues in Car / Handheld

- PTT – Barge-in
- Visual feedback/screen
  - → N-best lists for disambiguation in Navigation system
- System latencies
  - Data access (DVD)
  - Recognition
- Platform & accuracy restrictions vs Naturalness
  - Example: destination entry UI if max 3000 active entries in grammar?
- Fallback strategies
- Open vs. directed/constrained dialog
- External factors competing for attention:
  - Driver may pause due to traffic interrupts
  - Dialog state then?

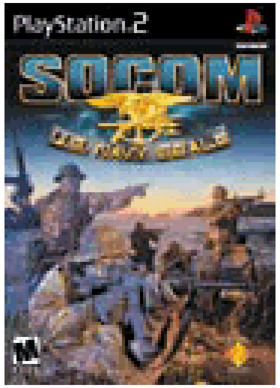
# Multimodality Today

- Car:
  - hands-free and possibly eye-free interface
  - → focus on speech, with (optional) visual feedback
  - Visual feedback: variable screen sizes
    - Car-kit: no screen
    - Navigation system: up to VGA color screen
  - → "light" multi-modality, with visual/haptic & speech I/O synchronized but no combined.
- Mobile:
  - Screen and buttons of phone → haptic & visual modalities
  - User will prefer faster/easier modality
  - Example: SMS dictation by voice, error correction via display/keyboard
  - → "light" multi-modality
- Game: part of game play design

# GPS Destination Entry Demo



# Video Game Demos






# State-of-the-Art vs. Cutting Edge

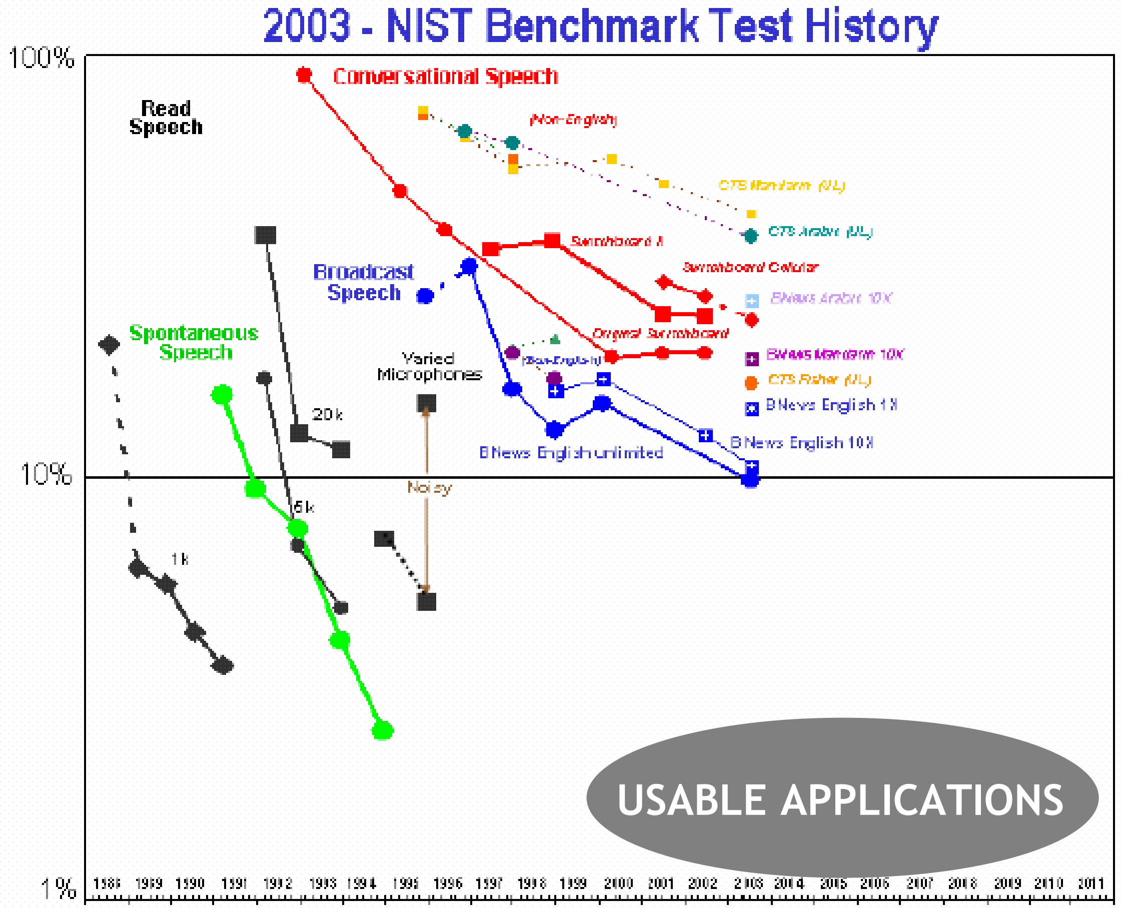



Figure 1 NIST Benchmark Test History

**A LOOK AT NIST'S BENCHMARK ASR TESTS: PAST, PRESENT, AND FUTURE** *David S. Pallett*

http://www.nist.gov/speech/history/pdf/NIST_benchmark_ASRtests_2003.pdf

# Embedded ASR Challenges

*Old Mantra:*

- *Improve accuracy*
- *Increase robustness*
- *Reduce footprint*

# Improve Accuracy

- How good are we today?
  - Example: Real CDD  WER in car is between 2—4% in at various driving conditions and across speakers today
- Where is the user acceptance threshold?
  - We are certainly not high above!
- Accuracy may limit complexity of task and constrain UI
- Must operate within platform constraints

# Improve Robustness

- Environment variability
  - Noise
  - Car in rainy weather, convertible!
  - Cellphones in public places
- HW variability
  - Microphone specs / actual production
- **User variability!**
  - Reduce « goats » percentage
    - Ideally, system should work for each and every user
  - Non-native speakers and accents
  - Multi-lingual systems
    - Address books
    - MP3 players  with songs in various languages
  - Emotions (stress while driving, games)
  - Children voices

# Reduce Footprint

- Why footprint?
  - Everything else being equal, footprint will drive BOM
  - Alternately, reducing footprint will allow to do more on given platform
- What footprint?
  → CPU and/or memory!
- Examples:
  - Navigation: larger item lists to simplify UI design
  - Handset: allow continuous dictation instead of discrete dictation
  - Game: add speech in the remaining 2% of resources available on console

# Conclusion

- Speech is back in embedded:
  - Expect to see more of it in your car, in your cellphone, in your living room (esp. if you have kids) in the coming 2 years
- But: seemingly simple applications remain challenging!
  - Not all basic problems are solved!
  - More research, better algorithms needed
  - Heavily coupled with engineering issues

# Questions?