Voice Search for Everyday Life

Perspectives on designing and scaling speech products

David Mitby, Lead PM Mobile

03.02.2009

Topics

Voice Search applications often use **large grammars or statistical language models**, which can be computationally demanding. At the same time, most Voice Search applications are **targeting high-volume use**. This session will address both technology adaptations and delivery options that can support this double burden **without unacceptable latency**.

The hardest part: getting "high-volume use"

Volume = users x speech usage

Using Voice to Get Information Very Appealing

How appealing is the idea of being able to simply use your voice to get the information you need?

- Very appealing 67%
- Somewhat appealing 23%

90%

- Neutral 7%
- Not very appealing 1%
- Not at all appealing 2%

The demand is there…

Case Study: Local Search on Sprint Instinct

- Users love it… **81%** repeat usage
- Optimized for voice… **82%** of queries use voice
- Efficient… **3** touches to find ‘Metropolitan Grill’
- Significantly outperforms ‘search / user’ benchmarks

Available on all Instinct devices today

Designing to Attract High-Volume Use

Get a distribution deal (carrier, automotive, telephony)

Designing to Attract High-Volume Use

Get a distribution deal (carrier, automotive, telephony)

Focus where other input fails

- **Difficult to type**

Touch devices: speech matters (the bar is lower)

Task	Minimum # of steps *iPhone*
Calling a contact	**5** touches + **1** scroll
Finding a business	**4** touches + **20+** typing the request
Playing a song	**3** touches + **1** scroll
Getting traffic	**3** touches + **1** scroll
Text message	**5** touches + **1** scroll Up to **160** typing the message

Designing to Attract High-Volume Use

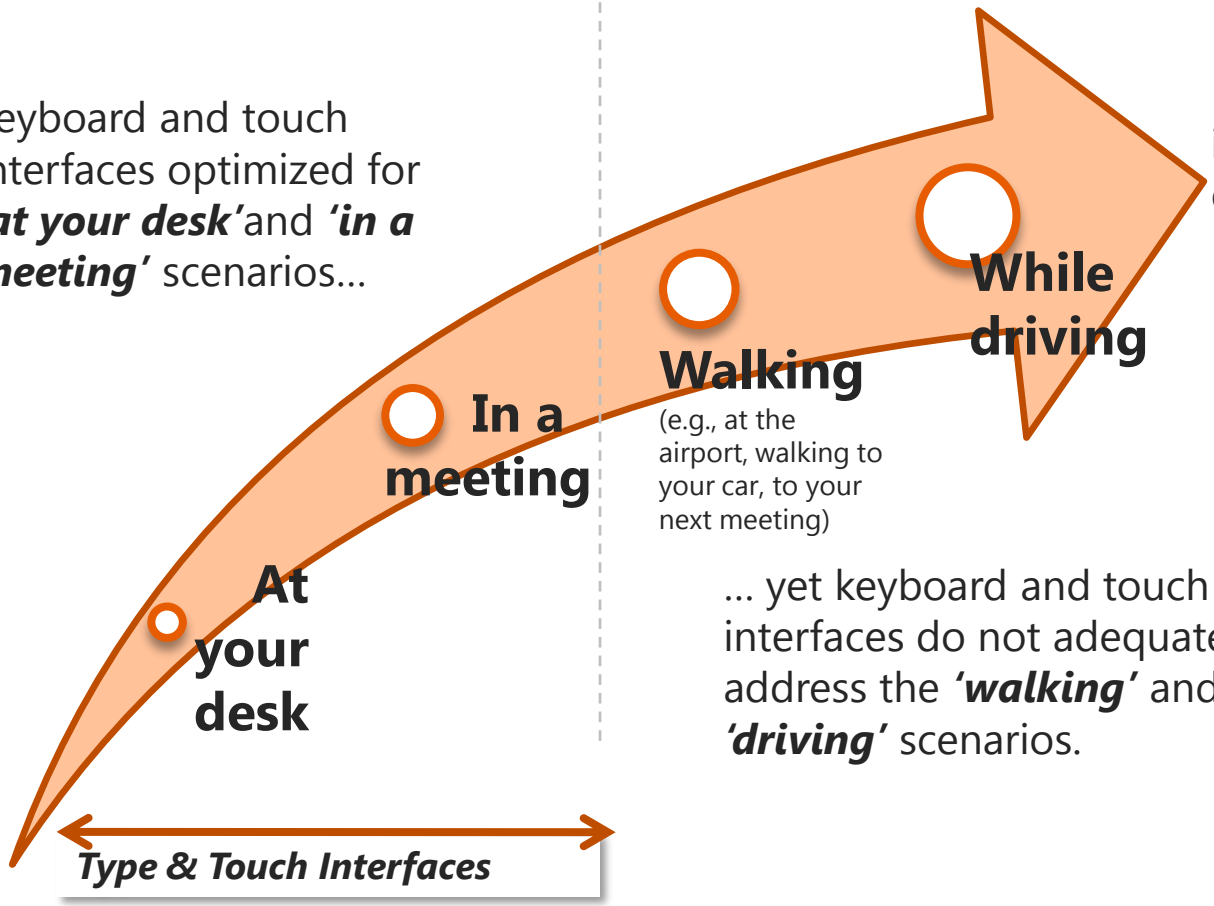
Get a distribution deal (carrier, automotive, telephony)

Focus where other input fails

- Difficult to type
- **"Distracted" situations**

Highest Value of Speech Use

keyboard and touch interfaces optimized for ***'at your desk'*** and ***'in a meeting'*** scenarios…



… yet keyboard and touch interfaces do not adequately address the ***'walking'*** and ***'driving'*** scenarios.

Context of use… distracted!

Designing for Distracted… old 411 vs. new:

Duration: 1:05

- Variable audio level
- Inconsistent voice talent and TTS
- Lots of user prompts
- Too many prompts
- Too long

Duration: 0:21

- **Personal:** Reduce number of steps (user inputs)
- **Seamless audio**: reduce jarring transitions, consistent audio throughout
- **Fast**: eliminate superfluous prompts and agent handling (automate)

Designing to Attract High-Volume Use

Get a distribution deal (carrier, automotive, telephony)

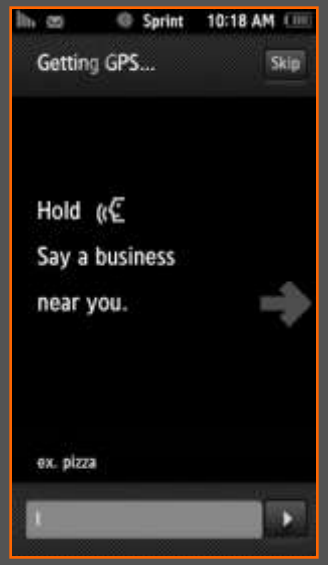
Focus where other input fails

- Difficult to type
- "Distracted" situations

Design with speech in mind

- **Teach! It's still new. And emphasize speech.**

User testing first-time use – and emphasizing speech



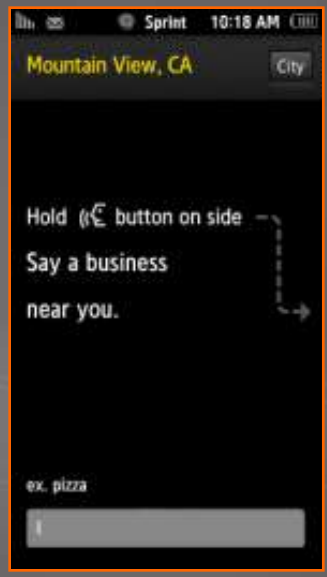
1

2

3

4

5



6

Designing to Attract High-Volume Use

Get a distribution deal (carrier, automotive, telephony)

Focus where other input fails

- Difficult to type
- "Distracted" situations

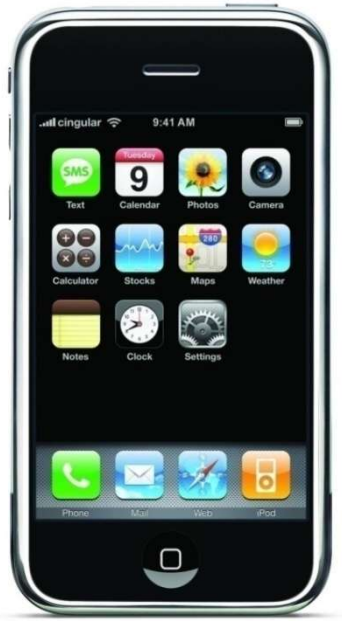
Design with speech in mind

- Teach! It's still new. And emphasize speech.
- **Enable quick access**

Make it easy to access – button, top-level tile…

Tellme on Sprint Instinct

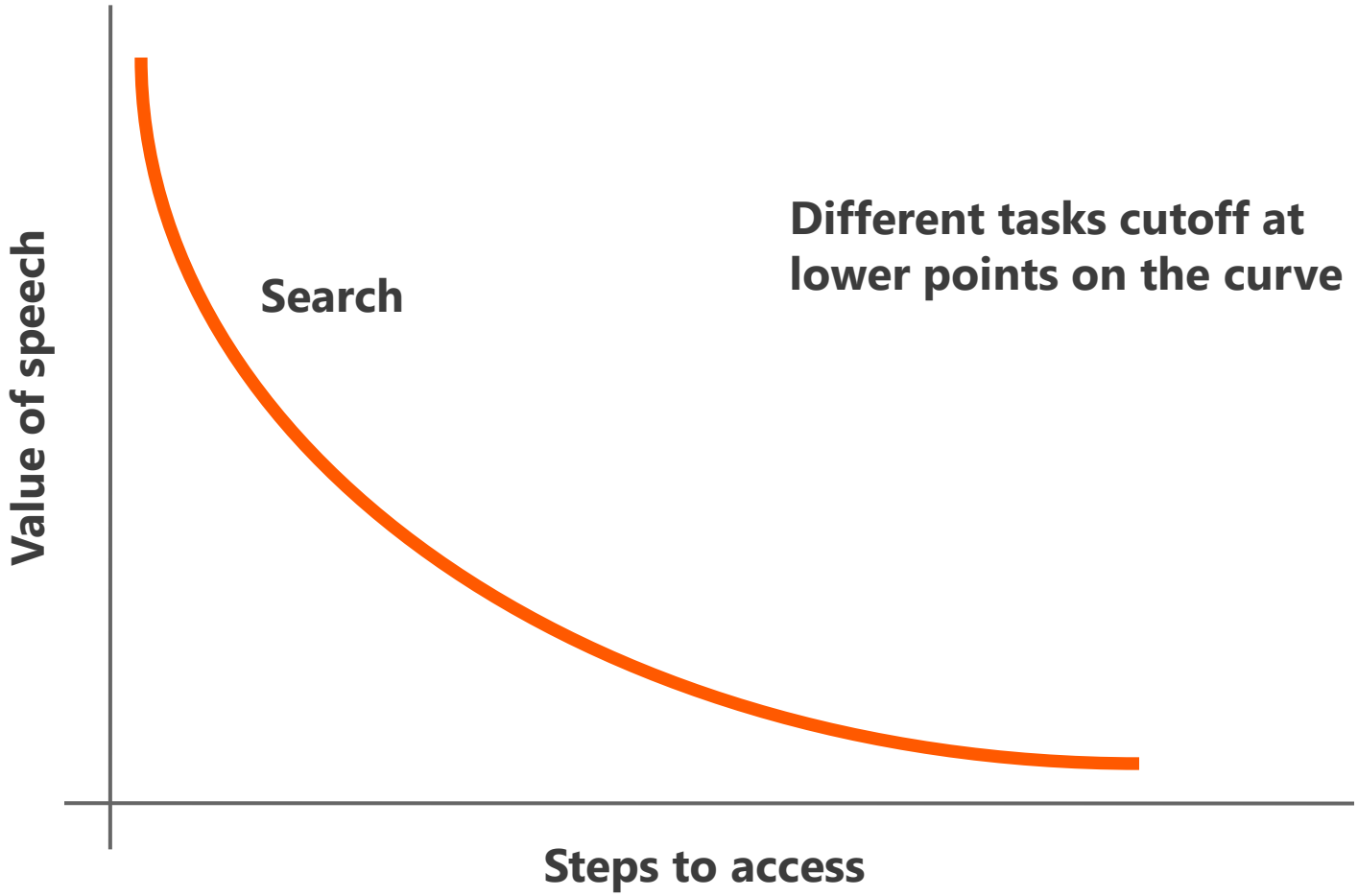
3 touches to find ‘Metropolitan Grill’



Search (without voice) on iPhone

23 touches to find ‘Metropolitan Grill’

Declining Value of Speech Based on Steps to Access



Designing to Attract High-Volume Use

Get a distribution deal (carrier, automotive, telephony)

Focus where other input fails

- Difficult to type
- "Distracted" situations

Design with speech in mind

- Teach! It's still new. And emphasize speech.
- Enable quick access
- **Minimize steps**

Minimizing Steps – Speech+Search Problem: Examples

Automatically halo

- "Home Depot" in Leawood, KS → Home Depot in Overland Park, KS

Recognizer n-best integration (weighted multi-input search)

- "Christine Salon" vs "Chris Bean Salon"
- "Pelican" vs "Palace Inn"

Highly tuned reco based on the data set

- Varies by type of data (local vs. web vs. VAD vs. combo)
- In general, great targeted synonym expansion, dictionaries, acoustic models, etc

Presentation Name Normalization to Minimize List Sizes

- Safe Way Market, Safeway Food & Drug, Safeway Stores → Safeway
- Pete's Coffee → Peet's Coffee

Designing to Attract High-Volume Use

Get a distribution deal (carrier, automotive, telephony)

Focus where other input fails

- Difficult to type
- "Distracted" situations

Design with speech in mind

- Teach! It's still new. And emphasize speech.
- Enable quick access
- Minimize steps
- **Voice != text search – deal with added complexities**

Voice != Text Search

Digits – “twenty four hour fitness”

Homophony – “sysco”/“cisco”, “citi bank”/“city bank”

Many more…

Now that you have users and usage…

Few parting thoughts on latency and scale

Topics

Voice Search applications often use large grammars or statistical language models, which can be computationally demanding. At the same time, most Voice Search applications are targeting high-volume use. This session will address both technology adaptations and delivery options that can support this double burden **without unacceptable latency**.

Thoughts on latency:

- Most latency in the network and handset, not backend (if done right) – so best experience will be on higher-end handsets and 3G networks
- Streaming audio end-to-end
- Compression: trade-off on accuracy vs. latency (though often limited options)
- Embedded vs. server-side vs. hybrid

Scaling the Network: Principles

Allow any machine to fail – build in units

Build in multiple datacenters to allow site-level failure

Plan for constant updates – successful recognition = constant tuning

Plan for downtime for software upgrades

Have a 24x7 NOC

In tradeoff between accuracy and more hardware, buy hardware

Redundant, geographically diverse data center facilities

Geographical Redundancy

- Geographic isolation from regional events and disasters
- Redundancy within & across data centers
- Multiple Tier-1 ISPs (multi-homing) at each data center

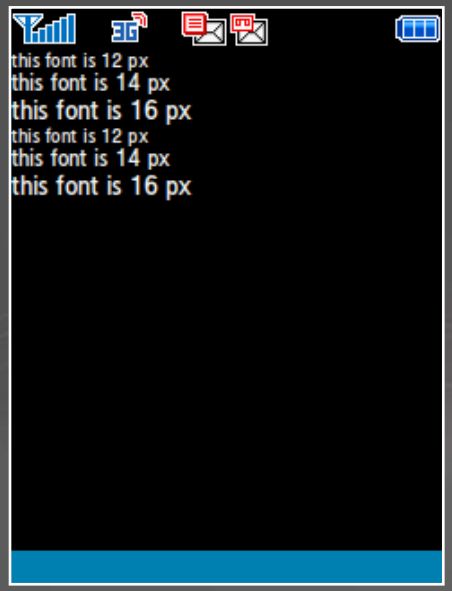
Robust Site-level Engineering

- Multi-tier redundancy & failover
- Distributed high-performance HTTP, grammar & audio caches
- Integrated network monitoring & mgmt
- Path-based instrumentation and analysis framework

99.995%

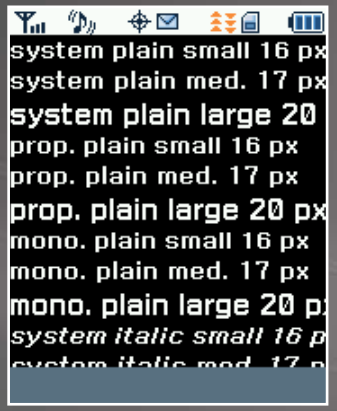
The *real* challenge to scale – *design*, not reco, due to form factors

240 x 320



Samsung A707

176 x 220



Samsung A920

128 x 160

Samsung X497