Voice and Multimodal Interaction in the Car

Garrett Weinberg

Nuance Communications, Inc.

AutomotiveUI , 17 October 2012

Quick poll

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Outline

- Understanding how the underlying technology works
- The big role that user interface design plays
  - Voice > manual? Not always…
- “Natural language” and what it means to different people
- Voice + X = Multimodality
- What the future might hold

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Outline

- Understanding how the underlying technology works
- The big role that user interface design plays
  - Voice > manual? Not always…
- “Natural language” and what it means to different people
- Voice + X = Multimodality
- What the future might hold

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Speech recognition fundamentals

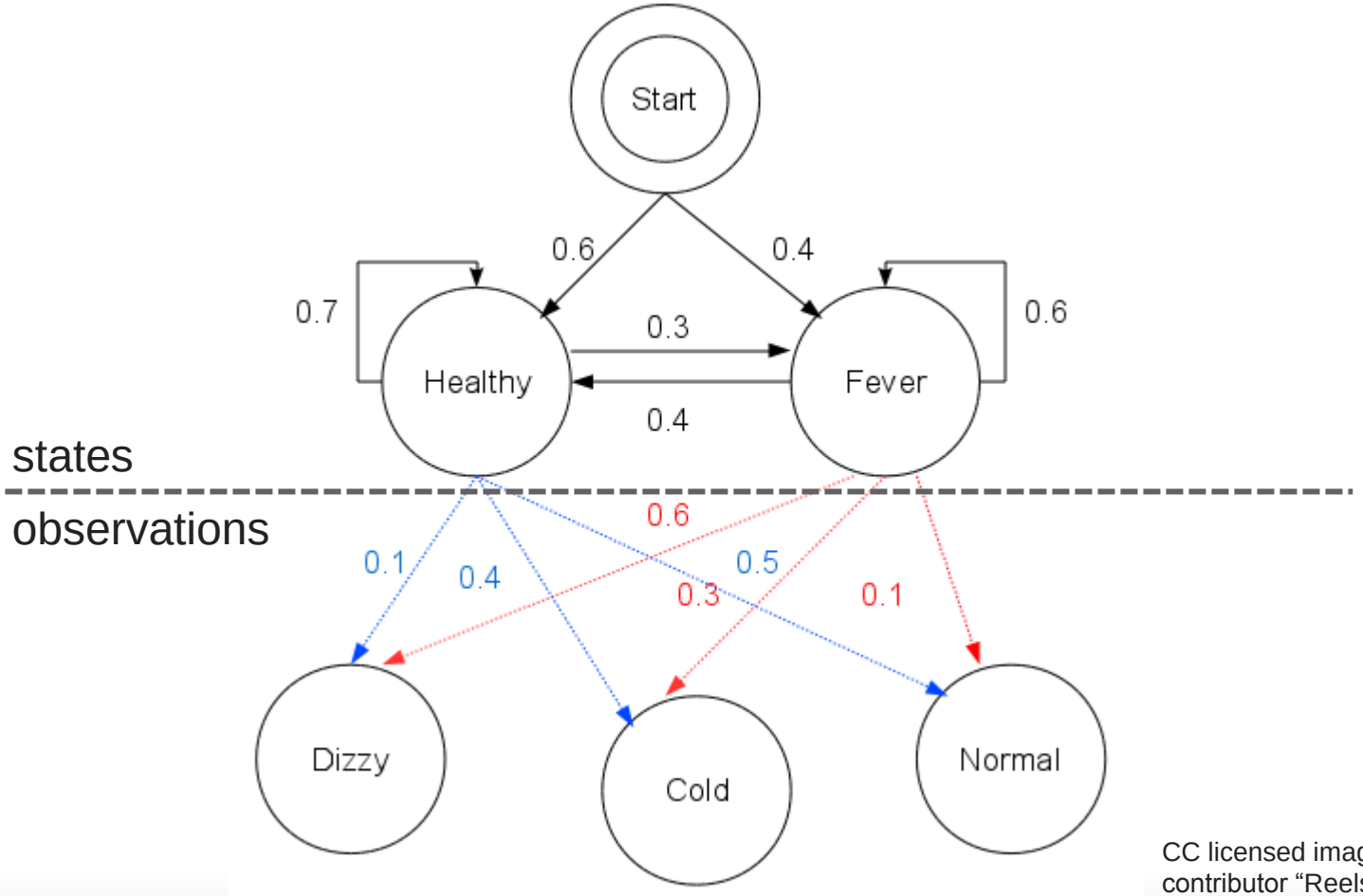
Standing on the shoulders of Markov and Viterbi

- Fundamentally, computers are best at calculating what *might* have been said with some (hopefully high) degree of likelihood. How does that happen?
- Markov processes (or chains) are those in which the next state depends only on the previous state and a fixed set of probabilities.
  - True about human speech? On the phoneme level, yes.
- Hidden Markov Models describe systems in which the Markov process cannot be directly observed, but must be guessed at based on a sequence of observations.
  - For automatic speech recognition (ASR), the observations are the series of acoustic signals recorded by the microphone and the “hidden cause” is the phoneme sequence that generated them
    - Candidate phoneme sequences → candidate words → candidate sentence

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Speech recognition fundamentals

Standing on the shoulders of Markov and Viterbi



CC licensed image by Wikipedia contributor “Reelsun”

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Speech recognition fundamentals

Standing on the shoulders of Markov and Viterbi

- Viterbi invented a dynamic programming-based algorithm for efficiently deducing the hidden states in an HMM
- Phoneme state transition and observation probabilities are fixed for a given language. These comprise the **acoustic models.**
- **Language models** help to constrain the Viterbi search in terms of memory and CPU.
  - Finite-state grammars:

    ```
    I (want | would like) [to order] [a | some] (pizza | beer)
    ```

  - Statistical language models:

    ```
    I want              0.0658
    I would             0.0433
    want to             0.1126
    want a              0.1410
    want bicycle        0.0001   (“want bicycle tires for Christmas”)
    ```


© 2002-2012 Nuance Communications, Inc. All rights reserved.

Speech recognition fundamentals

Resources

L. R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 2, 257-286.

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Speech synthesis fundamentals

The other side of the coin

- Fundamental problem: how to pronounce a word that is not known ahead of time?
  - "Turn left ahead" vs. "Turn left on Spooner St."
- Various techniques exist. They vary in:
  - Computational resource utilization
  - Intelligibility
  - Pleasantness

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Speech synthesis fundamentals

Major approaches

- **Formant synthesis** generates waveforms using models based on subword-level phonetics.
  - No recorded speech used
  - Fundamental frequency and consonant voicing varied over time to produce output signal
  - Small-footprint

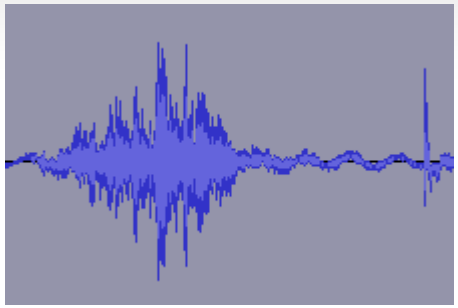
CC licensed image by Flickr user “farnea”

© 2002-2012 Nuance Communications, Inc. All rights reserved.

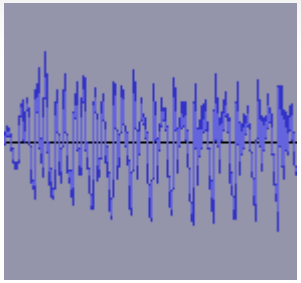
Speech synthesis fundamentals

Major approaches

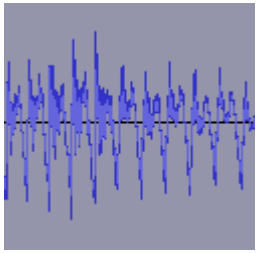
- **Concatenative** systems generate waveforms by splicing together small samples of real human speech.
  - Pitch and timbre of vowel sounds varies greatly according to their surrounding consonants
  - Therefore sample databases can get quite large (up to multiple GB)
  - Signal processing can smooth out artifacts at sample boundaries



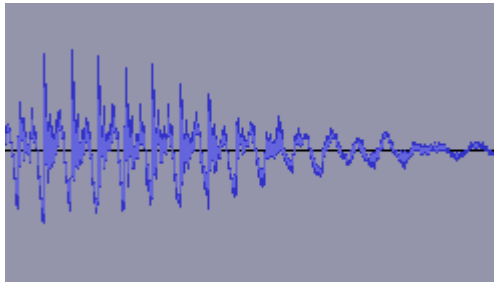
‘Sp’



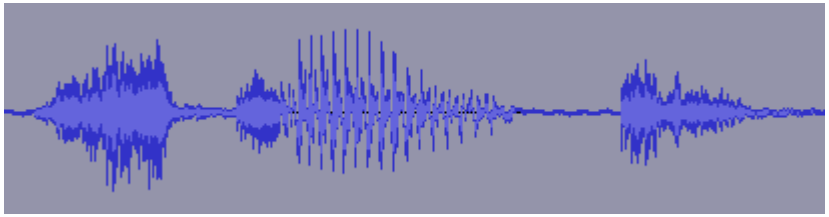
‘oo’ into alveolar



‘oo’ into ‘n’



‘n’ into terminal ‘er’



‘street’

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Speech synthesis fundamentals

Resources

Wikipedia article “Speech synthesis”
http://en.wikipedia.org/wiki/Speech_synthesis

A. J. Hunt and A. W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1. IEEE Computer Society, Washington, DC, USA, 373-376.

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Outline

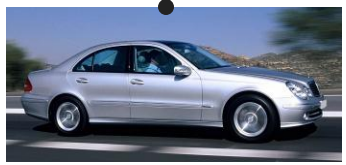
- Understanding how the underlying technology works
- The big role that user interface design plays
  - Voice > manual? Not always…
- “Natural language” and what it means to different people
- Voice + X = Multimodality
- What the future might hold

© 2002-2012 Nuance Communications, Inc. All rights reserved.

History Lesson

1996
Digit Dialing

2002
+ Command & control
+ 4-Ch Hands-free
+ Whole-word VDE



2004
+ Voice Destination Entry (VDE) by spelling
+2-Ch Handsfree

2006
+ 1st Korean

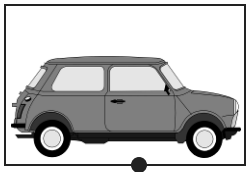
2008
+ Full Multimodal

2010
+ Japanese
+ 1-shot VDE

2011
+ Music Search

+ Multilingual Music

+ Apps
+ Dictation
+ Named POI
+ in-car comm.



© 2002-2012 Nuance Communications, Inc. All rights reserved.

Voice User Interfaces

It's not what you say, but how you say it

- Historically, car makers sometimes guilty of "feature-itis"
  - Highly competitive marketplace
  - TV commercial and showroom experiences paramount
- Can lead to the design decision: voice-enable everything!
  - Commands such as "Volume up," "Preset 1"
  - Decades of refinement in switches and knobs. If the tactile/haptic option is aesthetically pleasing and effective… is voice really better for such situations?
- Consider reserving voice for where it's most valuable
  - Searching, browsing and filtering

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Voice User Interfaces

It's not what you say, but how you say it

- What makes a good VUI?
- Something of a subjective question, but certain heuristics have evolved.
- System / mental model harmony:
  - When do I speak? (predictably timed tones, status indicators)
  - What can I say? (availability of commands)
  - When can I say it? (comprehensible system state)
  - How can I say it? (flexibility of formulation)
  - Can I get help? (e.g. "what can I say?" command)
  - In detail? (tutorial)

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Voice User Interfaces

It's not what you say, but how you say it

- More heuristics:
  - Does the system understand me? (ASR / intent accuracy)
  - Can I get out (or back)? (back / undo / cancel functions)
  - If the system isn't sure, does it let me choose? (confirmation behavior)
  - What if there are multiple "right" answers? (browsing support)
  - If I'm not sure (or busy), will it wait for me? (hesitation support)
  - What if the system makes a mistake? What if I do? (error recovery)

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Voice User Interfaces

Resources

General “rules of thumb:”

B. Schmidt-Nielsen, B. Harsham, B. Raj, and C. Forlines. 2008. Speech-Based UI Design for the Automobile. In Lumsden, J., ed., Handbook of Research on User Interface Design and Evaluation for Mobile Technology. NRC of Canada, Ottawa. 1, 15, 237-252.

The important role that ASR accuracy plays:

A. Kun, T. Paek and Z. Medenica. 2007. The Effect of Speech Interface Accuracy on Driving Performance. In Proc. of Interspeech.

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Evaluating Voice User Interfaces

Does it get the job done?

- Non-heuristic methods of evaluating a VUI:

  Usability:
  - Task time
  - Success rate (both at utterance and task level)
  - Satisfaction
  - Enjoyment

  Fitness for Purpose:
  - Task time (again)
  - Glance duration and frequency
  - Hand(s)-off-wheel time
  - Lane deviation and/or steering wheel angle variation
  - Acceleration and braking behavior
  - Following/leading distance maintenance
  - Reaction to stimuli
  - etc.

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Evaluating Voice User Interfaces

Resources

Good survey of pre-2006 work:

Barón, A. and P. Green. 2006. Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review. Technical Report UMTRI-2006-5

A/B Comparisons of manual/voice interfaces:

Maciej, J., and M. Vollrath. “Comparison of Manual vs. Speech-Based Interaction With In-vehicle Information System.” *Accident Analysis and Prevention.* 41 (2009): 924–930

Garay-Vega, L. et al. "Evaluation of Different Speech and Touch Interfaces to In-vehicle Music Retrieval Systems.” *Accident Analysis & Prevention*. 42.3 (May 2010).

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Evaluating Voice User Interfaces

Resources, cont.

Tools of the Trade:

Gärtner U. et al. Evaluation of Manual vs. Speech Input When Using a Driver Information System in Real Traffic. In Proceedings of The First International Driving Symposium on Human Factors in Driver Assessment (2002).

Harbluk, J. L. et al. Three Navigation systems with Three Tasks: Using the Lane-Change Test (LCT) to Assess Distraction Demand. In Proceedings of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design (2010).

Mahr, A. et al. The ConTRe (Continuous Tracking and Reaction) Task: A Flexible Approach for Assessing Driver Cognitive Workload with High Sensitivity. In Adjunct Proceedings of the Fourth International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI). Portsmouth, NH, 17 – 19 October 2012.

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Outline

- Understanding how the underlying technology works
- The big role that user interface design plays
  - Voice > manual? Not always…
- “Natural language” and what it means to different people
- Voice + X = Multimodality
- What the future might hold

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Natural Language Understanding

Don’t talk like a robot

- Finite-state grammars (FSGs) :

```
I (want | would like) [to order] [a | some] (pizza | beer)
```

- These work well only up to a point. Always a missing combination or synonym…
- Contemporary embedded platforms are capable of running statistical language models (SLMs) instead of or alongside FSGs.
  - Plus: always-on 3G/4G connections to powerful servers!

```
I want              0.0658
I would             0.0433
want to             0.1126
want a              0.1410
want bicycle        0.0001
```


© 2002-2012 Nuance Communications, Inc. All rights reserved.

Natural Language Understanding

Don’t talk like a robot

- The use of SLMs opens up more flexible syntax.
  - Politeness words
  - Helping verbs
  - Flavoring words
  - Synonyms
  - Contractions
  - etc.
- “Radio tune 93.7 FM” → “I’d like to listen to 93.7 on the FM dial please.”

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Natural Language Understanding

It depends on what you mean by “natural.”

- However, you don’t necessarily need a high-powered machine with a high-speed Internet connection to offer a “natural” user experience.
- A lot depends on the user’s perception of the system
  - How to shape this perception: delicate design question

- Sometimes users are quite satisfied or even *more* satisfied issuing shorter, keyword-based commands or queries

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Natural Language Understanding

Resources

Application of general NLU concepts and system components onto the automotive domain:

F. Weng et al. 2006. CHAT: A conversational helper for automotive tasks. In Proc. of INTERSPEECH.

How do people actually “naturally” formulate their commands?

U. Winter et al. 2010. Language pattern analysis for automotive natural language speech applications. In Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '10). ACM, New York, NY, USA, 34-41.

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Outline

- Understanding how the underlying technology works
- The big role that user interface design plays
  - Voice > manual? Not always…
- “Natural language” and what it means to different people
- Voice + X = Multimodality
- What the future might hold

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Multimodality

More than one way to skin the cat

- Traditional computer interfaces: visual output, keyboard input
- Multimodal interfaces: N outputs, M inputs
- Engage more than one sense at a time
  - Hearing (speech input, audio and speech output)
  - Touch (button, touchscreen, 2D & 3D gestural input; haptic output)
  - Sight (gaze-based input, one or more displays for output)
  - Smell and Taste? (some are working on it…)
- Also up-and-coming: Brain-Computer Interaction (BCI)

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Multimodality

Cornerstones

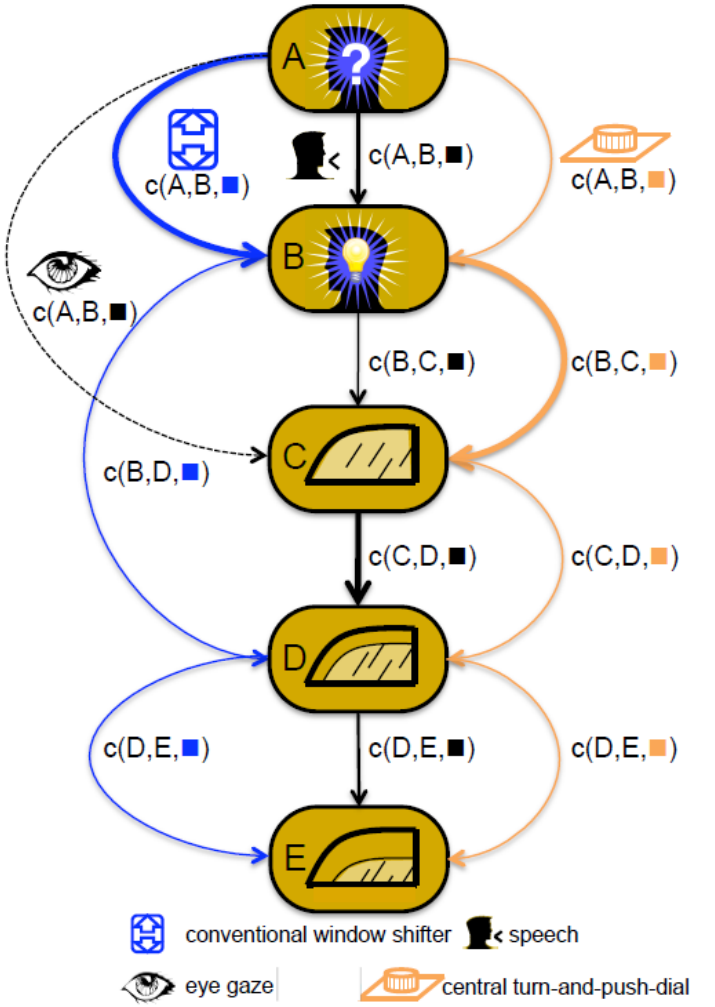
- Early example of multimodality (1980): famous “Put-that-there” demo from Richard A. Bolt of MIT.
- Sharon Oviatt synergized and expanded upon much of the theory. Here’s a simplified summary:
  - **Temporally cascading** modalities: input supplied in one modality informs or constrains another
  - **Redundant** modalities: you can choose your preferred modality at each step
  - **Fused** modalities: each modality is vital to the semantics of the overall action

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Multimodality

Not all modalities are well-suited to all parts of a task

- AB: ideation. Where is the window lever (or the multifunction knob)? What command do I use?
- BC: objectification. A command like "lower the window." Somewhere in the menu hierarchy… Or I could just look at the window.
- CD: execution. Easy enough. Here too. Do I say "a little bit?" How much is that?
- DE: repetition. Easy again. Here too. Now can I just say "more?"



(Taken from resource 2 on next page)

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Multimodality

Resources: theory

1. S. Oviatt. Multimodal interfaces. In A. Sears and J. A. Jacko, editors, *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, Second Edition, pages 413–432. CRC Press, 2 edition, Sept. 2007.

2. C. Müller and G. Weinberg. Multimodal input in the car, today and tomorrow. IEEE Multimedia, 18(1), 2011.

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Multimodality

Resources: practice

How temporal modality cascading can save time:

G. Weinberg et al. Contextual push-to-talk: shortening voice dialogs to improve driving performance. In Proceedings of the 12th international conference on Human computer interaction with mobile devices and services (MobileHCI), pages 113–122. ACM, 2010.

B. Pfleging et al. Multimodal Interaction in the Car: Combining Speech and Gestures on the Steering Wheel. In Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI). Portsmouth, NH, 17–19 October, 2012.

How modality fusion can add naturalness:

S. Castronovo et al. Multimodal dialog in the car: Combining speech and turn-and-push dial to control comfort functions. In Proceedings of Interspeech (2010), pages 510–513. ISCA, Makuhari, Japan, 26–30 September 2010.

© 2002-2012 Nuance Communications, Inc. All rights reserved.

Outline

- Understanding how the underlying technology works
- The big role that user interface design plays
  - Voice > manual? Not always…
- “Natural language” and what it means to different people
- Voice + X = Multimodality
- What the future might hold

© 2002-2012 Nuance Communications, Inc. All rights reserved.

The Future of In-Car Speech

Fundamental technology

- More functionality and more content at your fingertips
  - Faster embedded CPUs & more memory = more unconstrained Viterbi search space = more commands and formulations available at the top level of the application
  - Off-board and hybrid ASR

© 2002-2012 Nuance Communications, Inc. All rights reserved.

The Future of In-Car Speech

The bigger picture

- Voice UIs permeate lower- and mid-range cars
  - Apple's Siri helped make voice mainstream, but Siri's in-car mode is unimodal (blank screen)
- Car makers will start to take a more holistic, multimodal view of their HMI
  - Can they out-Apple Apple?
- What do **you** think?

© 2002-2012 Nuance Communications, Inc. All rights reserved.