novauris

Speak, *Find.* ...*EASY!*

Single-Input Voice UI for Navigation Systems

Yoon Kim
Novauris Technologies

Benjamin Ao
Alpine Electronics Research of America

2010 Mobile Voice Conference

2010-05-05

Company Intros

novauris

- Founded in 2002 by key members of Dragon Systems
- Core ASR technology for search and access of very large lists
- Server, embedded and client-server hybrid offerings
- Powering the voice search for Verizon Get It Now Search
- Deployments in US and Japan

ALPINE

- Founded in 1978 as a consolidated subsidiary of Alps Electric Co., Ltd.
- Dubbed a “mobile media solutions” company, Alpine specializes in integrating digital entertainment, driver-assist and navigation in mobile electronic devices for the automotive
- Pioneer in bringing voice technology into the automotive environment
- Alpine technology is deployed in OEM and AFT products world-wide

Outline

- VUI for mobile and navigation applications
- "Natural" voice interfaces
- Destination entry for navigation systems
- Single-input vs multi-input: which is better?
- ASR in car noise
- Experiment: single-input address recognition (by Alpine Electronics Research)
- Summary

VUI for Mobile and Navigation

- **Observations from real mobile voice search users**
  - Users want relevant information quickly: *"I want it now, and it better be what I had in mind!"*
  - Users are lazy: *"No pain… no pain!"*
  - Users don't want to converse with machines
  - Users adopt speech input as their preferred mode… but only AFTER building trust in the system VUI

- **In-car environments: Any different?**
  - Operating a "mobile device" that is bigger and faster than you! ☺
  - The above observations about users are still valid
  - Consider user cognitive load and anxiety level in coping with HMI

"Natural Language" Interface: Lessons Learned

- **"Natural speech" is not a single specific speaking style**
  - it is the most appropriate style for the situation and task
  - it's the style that speakers adopt when not given specific instructions

- **In most voice input modes, "natural" means terse**
  - speakers naturally use direct language with the fewest possible number of words
- **It's essential to anticipate the "natural" speaking style for the particular application**
  - Because of the trade-off (more flexibility comes at a price of lower accuracy and increased computational load)

- **Automotive: Hands-free operation and cognitive load**
  - Users may prefer reliability/speed over input flexibility

Destination Entry for Navigation

- **Navigation** provides convenient, reliable & safe guidance
  - especially to new, unfamiliar destinations
  - ***BUT*** destination entry is currently not convenient and not quick
- **Effective voice-enabled destination entry system:**
  - Accuracy; speed; easy launch & access; intuitive error correction
- **Addresses**
  - Highly structured and entire input is usually lengthy
  - Order of input varies by region (e.g. US vs Japan)
- **POIs**
  - Mainly unstructured except when tied with location (e.g. city, state)
  - Need to consider variants and aliases when spoken

Single Input vs Multiple Input

Please speak the Street address:

"555 North Point St. San Francisco, CA"

Please speak the state:	"California"
Please speak the city:	"San Francisco"
Please speak the street name:	"North Point St"
Please speak the street number:	"555"

Quicker, easier, safer, and more "natural"!

Single-Input vs Multi-Input

- VUI design for in-car navigation: Accuracy, speed and noise robustness of ASR directly affects task completion rates

Single-input entry

- Quick, safe and convenient, if ASR performance is reliable
- Error correction can be tricky
- Ideal for power users
- Lengthy inputs could be a challenge to some users

Multi-input w/ confirmation

- Tedious and slow
- Unnatural if order is reversed
- Error correction is built in
- Suitable for novice users

- We could consider a mix of single- and multi-input modes
  - Start with single-input and fall back to multi-input for error correction
  - More examples in the next slide!

Some Examples of Variants

Speak the street number and name:

"555 North Point"

Please speak the city and state:

"San Francisco, California"

555 North Point St, San Francisco, CA. Is that right?

"Yes"

Please speak the Street address:

"555 North Point"

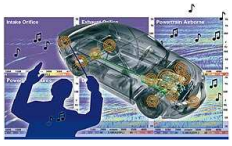
555 North Point St, San Francisco, CA. Is that right?

"Yes"

Current city/state optional; Location based constraints

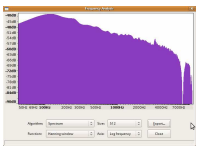
Evaluating ASR in Noise: Some Pitfalls

- It's not reasonable to assume that noise is always steady
  - the most disruptive noises are time-varying
- "SNR" is not well defined for ASR
  - the effect of noise on ASR depends on
    - its distribution over the spectrum
    - the spectrum range and warping used by the ASR
- An increase of N dB in the noise does not make the SNR N dB worse
  - People speak louder in noise
    - the change in SNR is closer to 0.5N dB

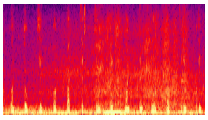


ASR in Car Noise

- Car noise is concentrated at low frequencies:
  <300 Hz, peaking below 100 Hz
  - so ASR much better with 5kHz range rather than telephone band



- The noise is usually steady over an utterance (except in the case of background music or speech from other passengers)
- However, with windows or the sunroof open,
  - the noise has disruptive high-freq components
  - and it's often not steady
    - passing traffic, heavy vehicles braking, horns…



- Novauris finds its single-shot address entry works very well with windows closed at high- and low-speeds
  - though somewhat less reliably with windows open
  - tests by several third parties confirm both observations

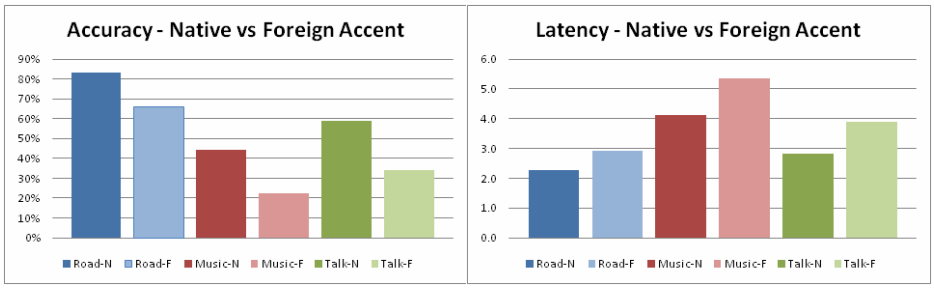
Alpine Testing of Novauris ASR: Single-Input US Address Recognition

- Task: NSCS (number-street-city-state)
  - Over 5M unique street names + "natural" house numbers (1~130,000)
  - Optional directional prefix/suffix, street type
- Test set
  - 30 speakers living in the US (10 native, 20 foreign born), 102 addresses each
  - 3,060 utterances recorded in studio, then digitally mixed with noise (road, music, talking) at different SNR levels, resulting in a total of 48,960 wave files
- System requirements and computational resources
  - Batch testing carried out on a desktop PC with 2.8GHz CPU
  - Peak CPU consumption: 50%; Peak memory consumption: 51MB

Note: Novauris embedded NSCS runs on a 400MHz+ CPU with <8MB RAM

Results: Single-Input US Address Entry

- Novauris system does well in road noise, but not so well for background music and talk
- Handles the various address input flexibilities well
- Appears to be ready for deployment in various voice-operated automotive navigation systems



Summary

- Users want VUI for automotive applications that is
  - quick, reliable, convenient, requiring minimum effort
- "Natural speech": most appropriate mode of speaking given the situation and task
- Navigation destination entry is a critical pain point for users
- Single-input entry: quicker, safer and natural, BUT needs to be reliable
- It is important to understand the characteristics of various types of car noises and its particular effects on the ASR performance
- Tests of our NSCS system show that in steady noise, native users may find our single-input address entry capability useful

Toward a Desirable Voice User Interface

Benjamin Ao
Alpine Electronics Research of America

Friday, April 23 (11:20am – 12:30pm)

Track 3: Speech in Consumer Products (Larkspur)

2010-05-05