Elderly Perception of speech from a computer

Maxine Eskenazi, Alan W Black and Reid Simmons
max@cs.cmu.edu, awb@cs.cmu.edu, reids@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

Background

Observation:
"*Elderly people can't understand computer speech.*"

Experiment:
– How can we make computer speech easier to follow

□ Two directions:
  – make voice more natural (current CMU research)
  – ***make delivery better for understanding

Goal:
– How can speech be modified to make it easier to understand

CMU's Vikia Robot

Listening Experiment

- Cueing Conditions

  V : voice alone

  VL : voice plus lip sync'ed head

  VM : voice plus robot moving, no lip sync

  VML : voice plus robot moving plus lip sync'ed head

Natural voice (not synthetic)

- 4 words pairs
  - common bi-grams:
  - *rose bowl*, *rose bush*
  - *holiday season*, *holiday shopping*
- 4 times
  - more constrained:
  - 4:28, 8:37, 11:52, 1:49

Subjects

- 23 subjects:
  - 8 male, 15 female
- average age 71.8 (std 6.4)
  - male 71.5/7.8
  - female 72.00/5.8
- from CMU Life Long Learning Program
  - High school and college educated
  - living in Pittsburgh

Results

Condition	Num	Words	Times
V	6	75%	100%
VL	5	100%	100%
VM	5	95%	100%
VML	6	100%	100%

Two people got 3/4 word pairs wrong in voice only
One person got 1/4 word paits wrong in voice plus move

Earlier Telephone Experiment

Synthetic and **Natural** Voice over Telephone

- Simple Computer Telephony Platform:
  – inexpensive LineJack on Linux box
- Subjects call and select session number:
  – defines the order they hear
- Listen and press key to continue

Subjects taken CMU's "Homecoming":
– elder (mostly aged 60+)
– mobile (can visit Pittsburgh)
– well educated (CMU graduates)

67 subjects ranging from 20s-80s
(some CMU staff and students)

Speech conditions

4 types of speech

- **NN**
  - natural spoken utterance
- **NS**
  - natural spoken utterance after
  - being told listener couldn't hear
- **SN**
  - synthesized diphone voice
- **SS**
  - synthesized diphone voice with
  - natural F0 and durations from NS

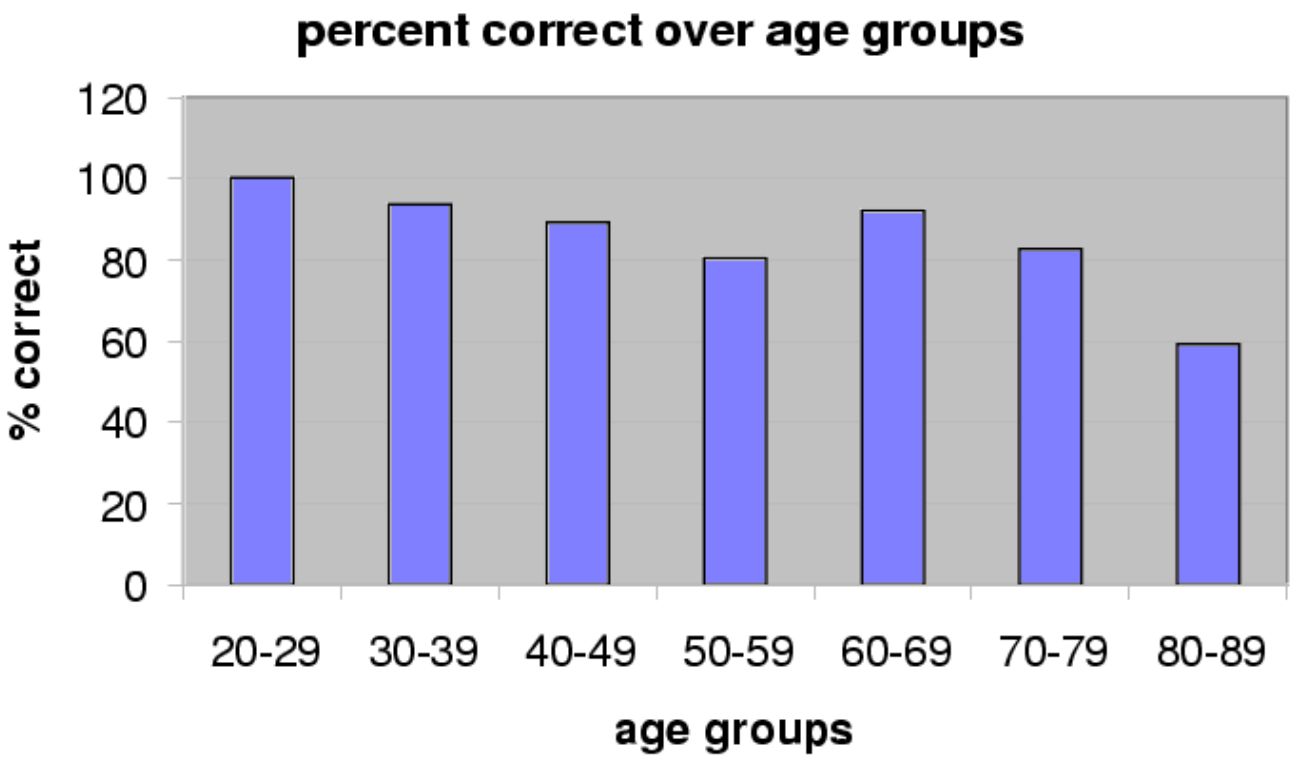
All voices female US speaker (synthesized voice based on same speaker).

Speech examples

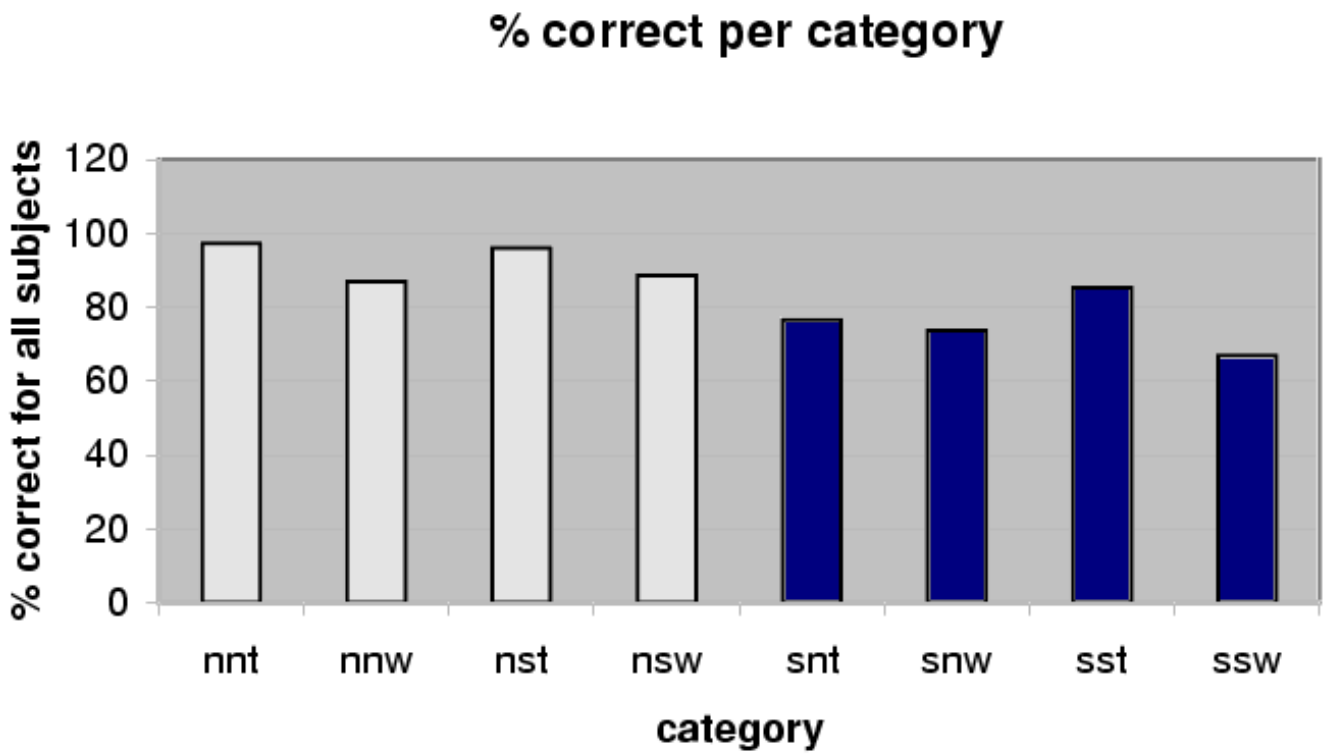
4 pairs of simple sentences:

- **Times**: constrained, allow listener to adapt to voice
  Please write down the following time ...
- **Words**: common bigrams,
  Please write down the following words ...
  – *holiday shopping*, *holiday season*,
  – *general motors*, *general manager*, ...
- Each pair in same voice
- Four voices in total
- 8 sentences in different orders

% Correct over age groups



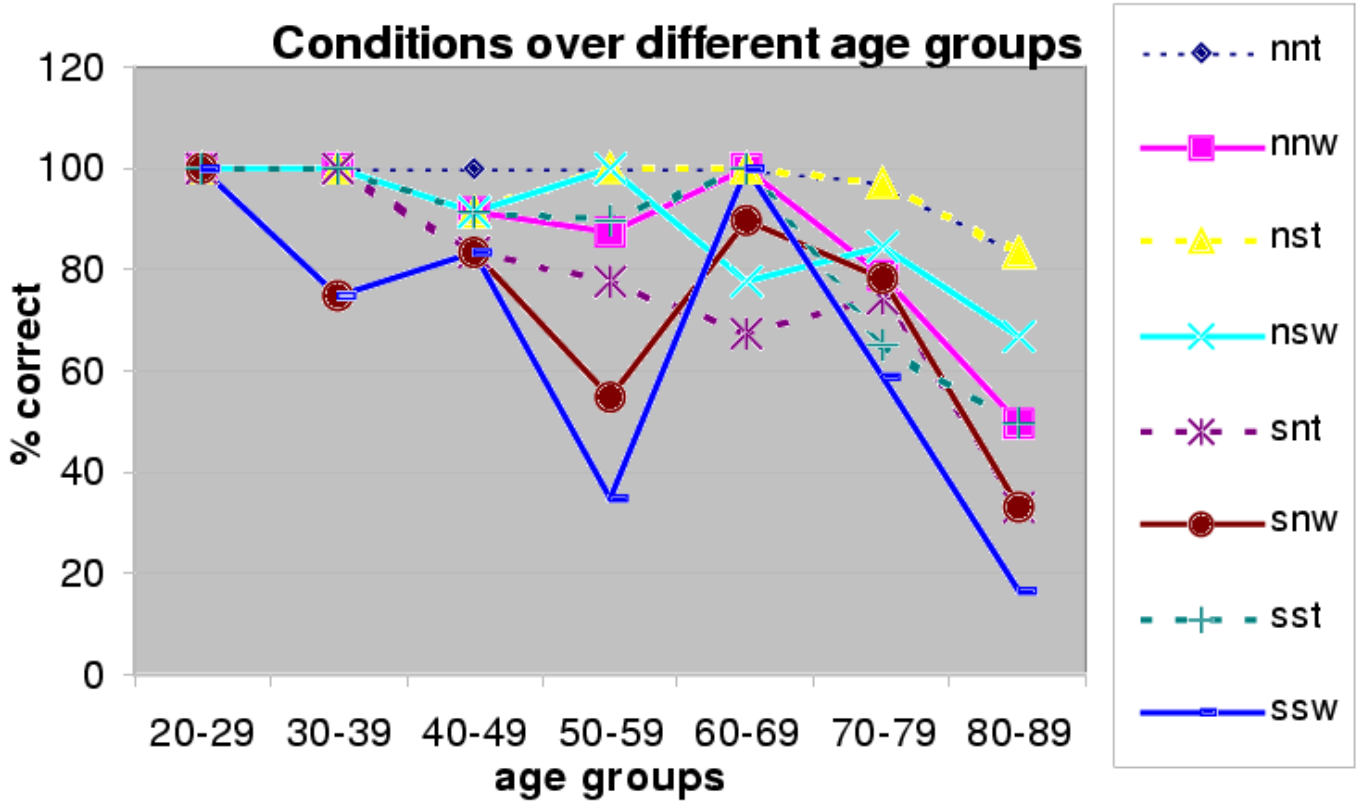
% Correct over category



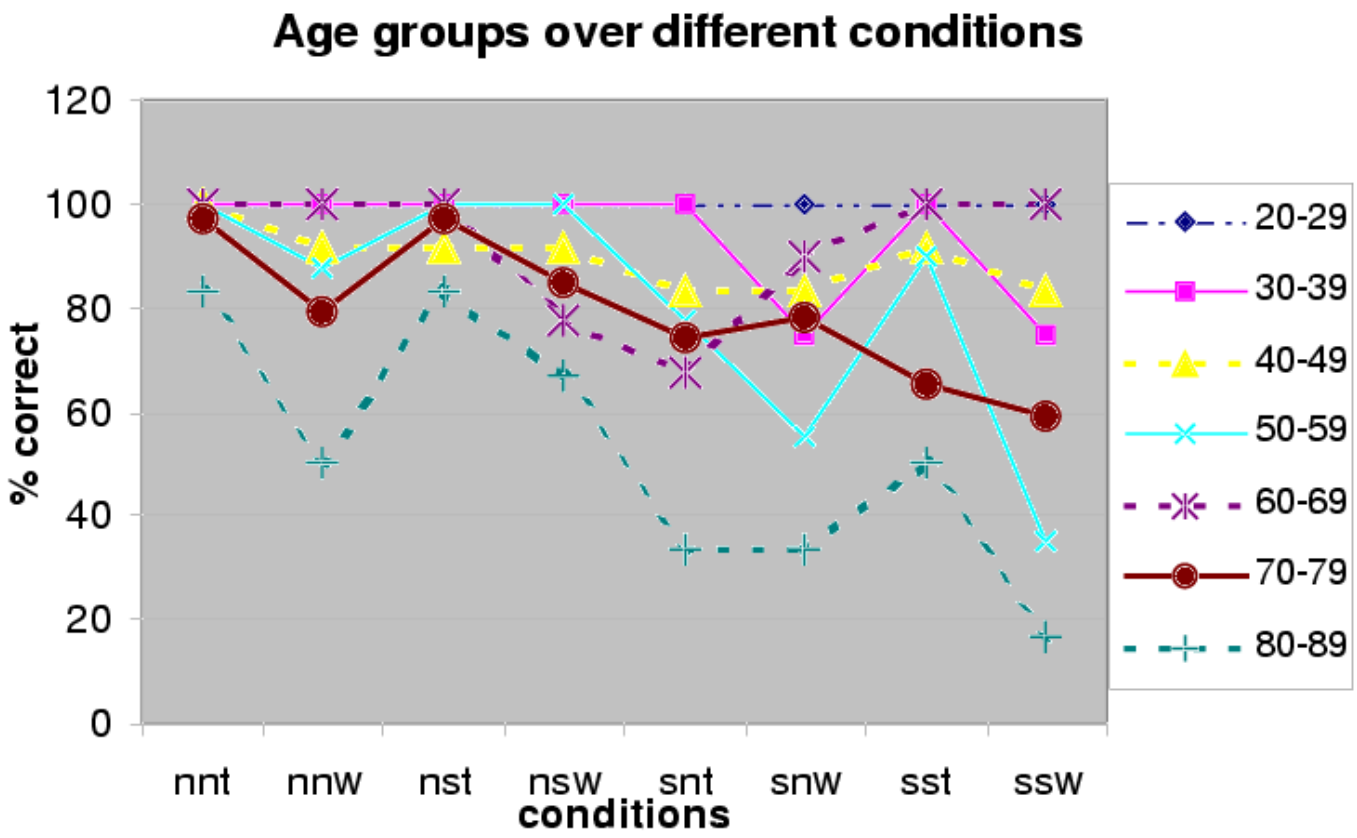
NN = natural speech
NS = natural clear speech
SN = synthetic diphone
SS = synthetic diphone plus natural F0

t = time
w = words

Conditions over age groups



Age groups over conditions



Conclusions and Future

- Earlier telephone study
  - understanding gets worse with aging
  - natural speech is better than synthetic
- Current experiment:
  - Voice alone is hard
  - Lip syncing helps (more than movement ?)
- Future experiments:
  - Confirm factors that aid understanding
  - Deploy in larger application for testing