Making Machines Talk and Listen - Why It Is Still Interesting

B. H. Juang

Georgia Institute of Technology

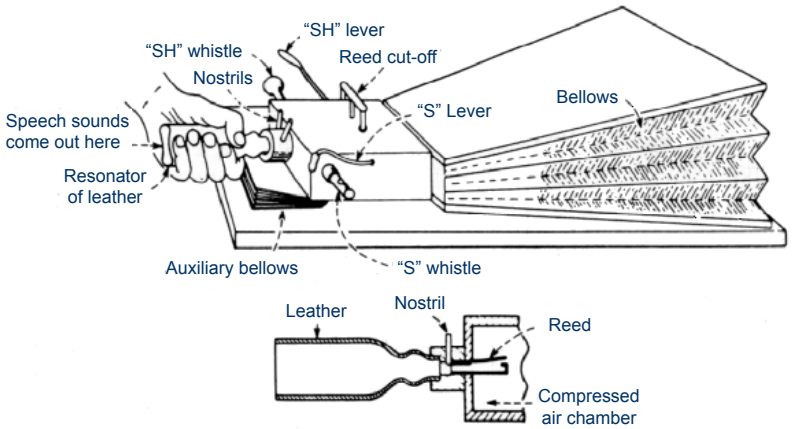
juang@ece.gatech.edu

Speech Processing Research

- Relevant History
- Technology Outlook
- Research Opportunities

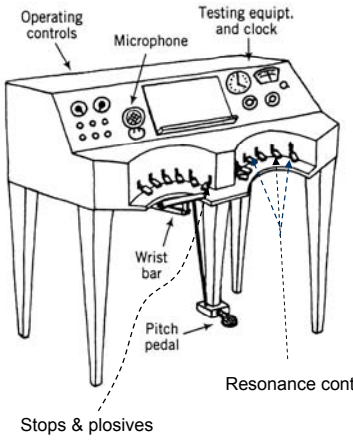
Speaking Machine by Wheatstone

Mimicry - Mimetic Signal Processing Model



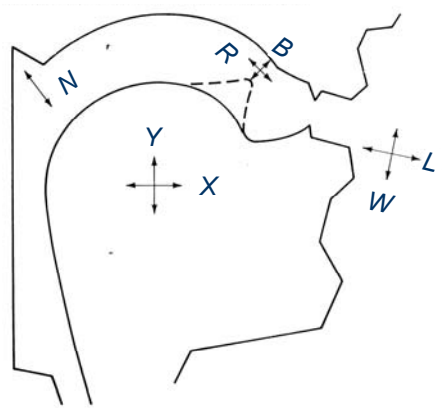
The VODER by Dudley & Later Improvement

Electrical



Articulatory Model & Synthesis

Coker's Computational Articulatory Model

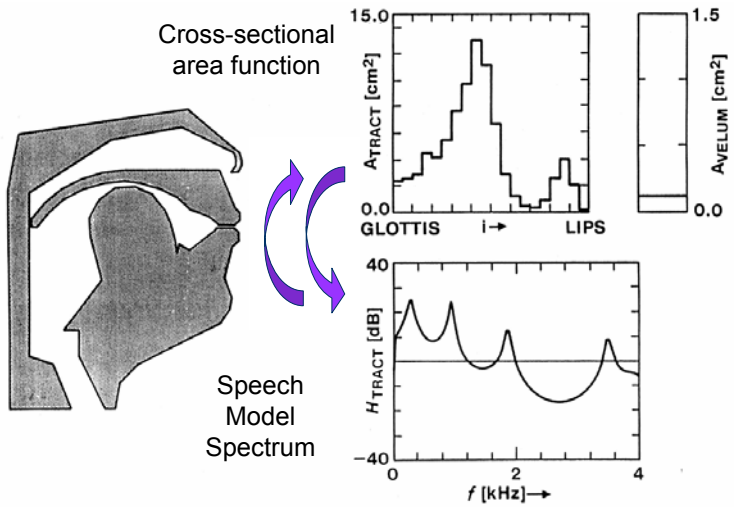


Articulatory Model Parameters

- lip *opening* – ***W***
- lip protrusion length – ***L***
- tongue body height and length – ***Y, X***
- velar closure – ***K***
- tongue tip height and length – ***B, R***
- jaw raising (dependent parameter)
- velum opening – ***N***

Georgia Institute of Technology

Vocal Tract Analysis with Linear Prediction



Georgia Institute of Technology

Continuing Evolution (1959-1987)

- Haskins, 1959
- KTH – Stockholm, 1962
- Bell Labs, 1973
- MIT, 1976
- MIT-talk, 1979
- Speak 'N spell, 1980
- Bell Labs, 1985
- Dec talk, 1987

Georgia Institute of Technology

Paradigm Shift in Speech Synthesis

- Parametric TTS has become highly intelligible but still extremely unnatural;
- **Parameterization amounts to data reduction, causing loss of subtle cues on naturalness;**
- Rapid growth in computation and memory diminishes the need in parameterization; an 80G HD can store 11,100 hours of good quality speech!
- New concatenative systems use segments of **recorded speech** to ensure better quality sound than any synthesized one with remaining challenge in **control**.

2.5" hard disc, 64.8 GB/in², 0.95cm (h) 80G < $200

Georgia Institute of Technology

Other Modern Commercial Systems

- Lucent
- AcuVoice
- Festival
- L&H RealSpeak
- SpeechWorks
- AT&T
- Cselt (Actor) – Italian

Georgia Institute of Technology

Text-to-Speech vs. Reading: Email Example

From: Marilyn Walker <walker@research.att.com>
To: David Ross <davidross@home.com>
Subject: Re: Today's Meeting
Date: Tuesday, December 01, 1998 4:25 PM

4:30 is fine for me. See you at the meeting.

Marilyn
-----Original Message-----

Being able to turn text into sound is NOT sufficient for many applications.

Thanks,
david ross

Conventional TTS

TTS with reading capability

Georgia Institute of Technology

A Historic Photo

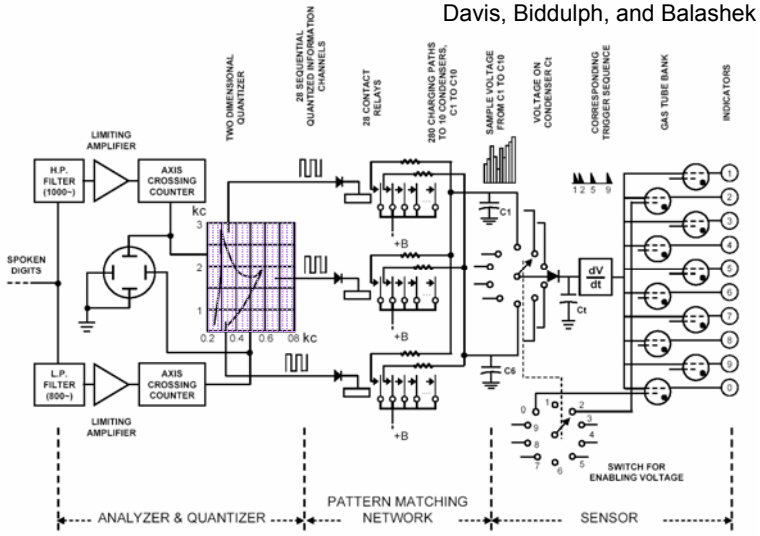
An early 20th century transcribing pool at Sears, Roebuck and Co.

courtesy www.recording-history.org

Georgia Institute of Technology

Early Digit Recognizer at Bell Labs - 1952

Davis, Biddulph, and Balashek

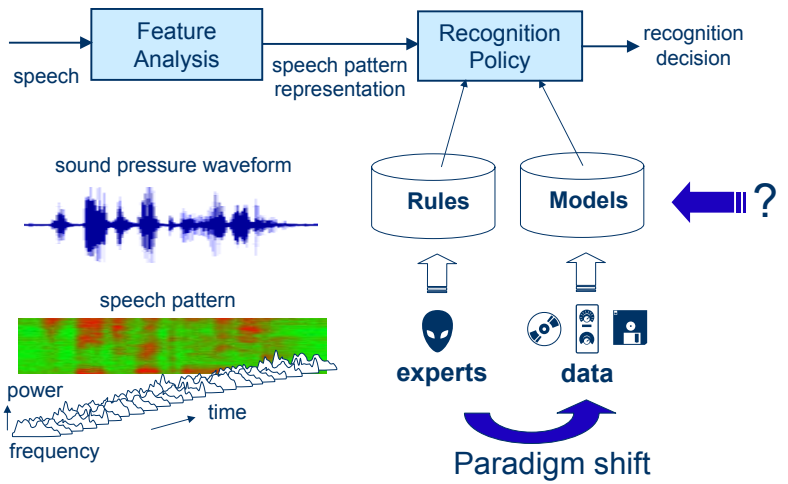


Georgia Institute of Technology

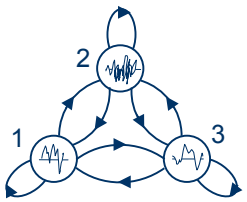
Regularity and Variability in Human Speech

- Regularity (**Structure** in Speech)
  - Linguistic hierarchy exists for any language: phonemes, lexicon, regular expressions, …
  - Speakers are taught to speak according to the convention of a language
- Variability (**Randomness** in Speech)
  - Casual use of words, improper pronunciation & word grouping, completeness in sentential structure or usual disfluency (um, repair) – ***to err is human***
  - Unregulated speaking pace & breath grouping relaxed pronunciation – ***to rest is human***

Speech Recognition Techniques



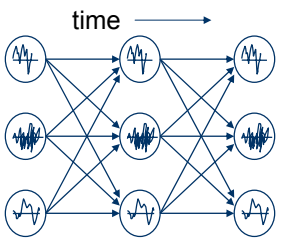
Hidden Markov Models



$$P(X \mid W, \lambda_x) = \sum_{\mathbf{q}} P(X, \mathbf{q} \mid W, \lambda_x)$$

$$P(X, \mathbf{q} \mid W, \lambda_x) = a_0 \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(x_t)$$

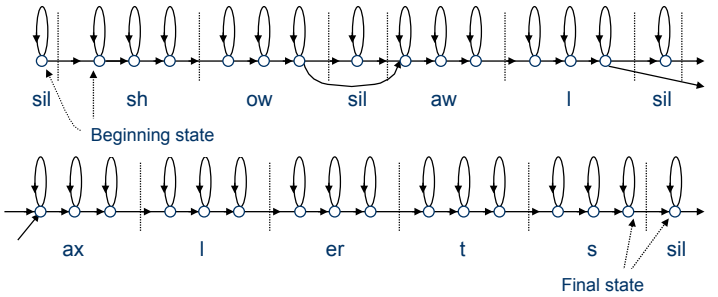
$$X = (x_1, x_2, \cdots, x_T) \qquad \mathbf{q} = (q_0, q_1, q_2, \cdots, q_T)$$



- Each state represents a process of measurable observations.
- Inter-process transition is governed by a finite state Markov chain.
- Processes are stochastic and individual observations do not immediately identify the state.

Embedding Acoustics in Finite-State Graphs

Composite Finite State Phone Network



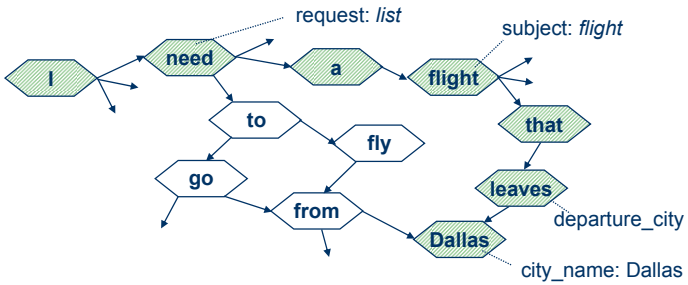
context-dependent phoneme model examples:
ϕ-**sh**-ow, ϕ-**ax**-l, ax-**l**-er, l-**er**-t

Language Models

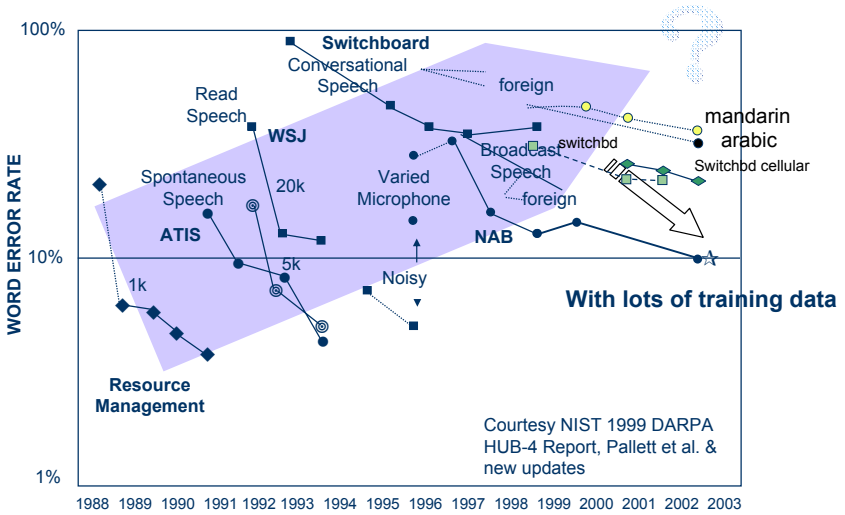
$$P(X,W \mid \Lambda) = P(W \mid \lambda_w) P(X \mid W, \lambda_x)$$

- N-gram: $P(W \mid \lambda_w) = \prod_{i=1}^{L} P(w_i \mid w_{i-1}, \cdots, w_{i-N+1})$
- General finite state: $P(W \mid \lambda_w) = \prod_{i=1}^{L} P(w_i \mid \lambda_w, H_i)$

  where H_i is the state of history for word w_i
- Domain dependency:

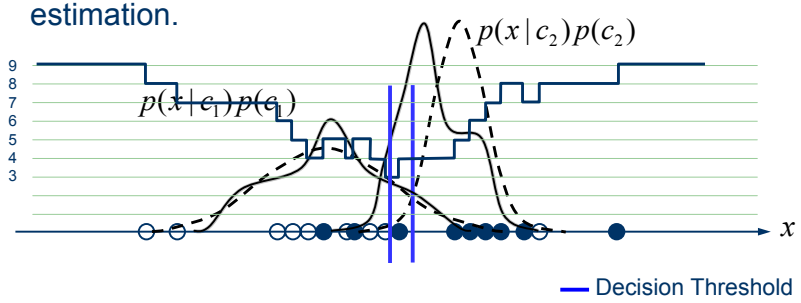


DARPA Speech-to-Text Benchmark Tests



Statistical Pattern Recognition?

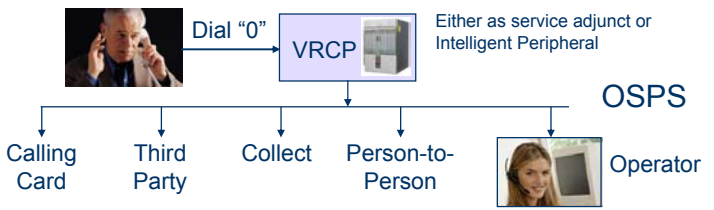
- Bayes decision needs knowledge of *a posteriori* probabilities ➔ conventional problem of distribution estimation.



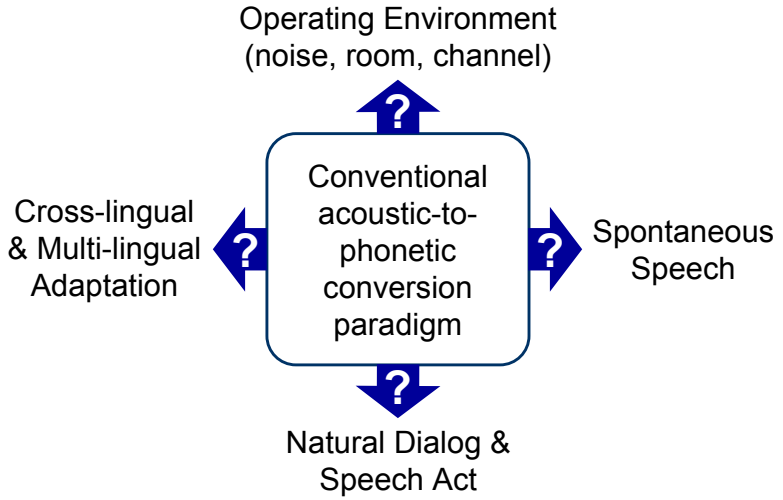
- Paradigm shift – look at the "error function" as an objective for optimization.
- Minimum error discriminative training leads to much improved recognition accuracy.

One Success Story – Call Automation

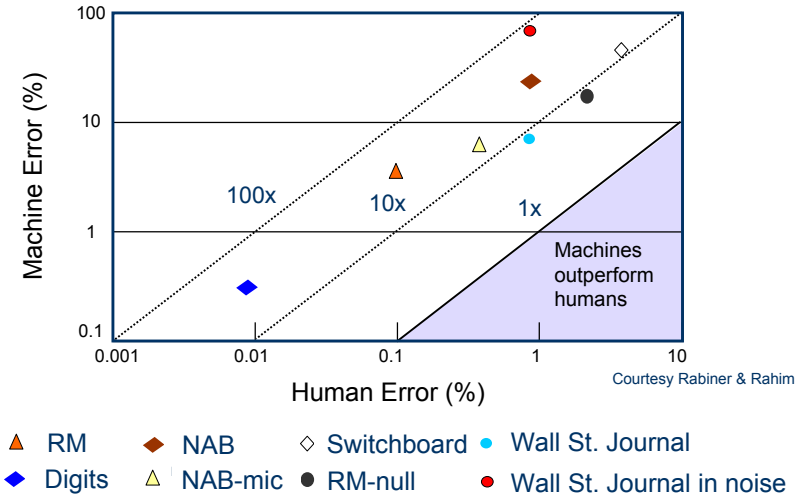
- Telephone companies used Operator Support Position System to handle operator calls, a precursor to call center
- Automatic speech recognition w/ **hidden Markov models** reached maturity by late 80s for some telecom applications
- Voice Recognition Call Processing (VRCP) by spotting five essential keywords handles billions of 0+ calls annually.



Current/Overall Technology Assessment



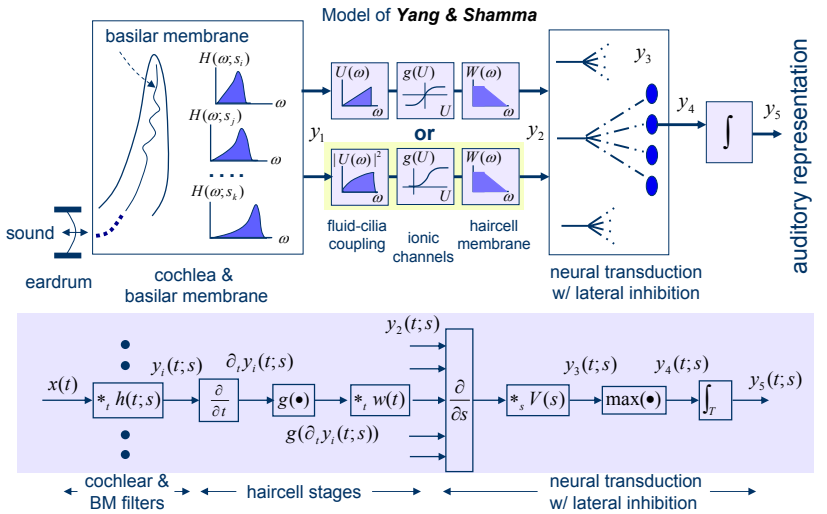
Human Still Outperforms Machines



Moving Forward

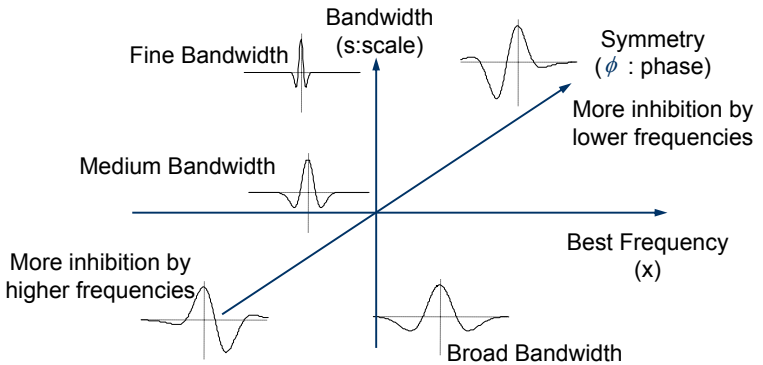
- Learn from human – why human does better?
  - Biological system models
  - Detection-based paradigm
- Address naturalness in human-machine interactions
  - Speech act and hierarchy; semantics
  - Dialog management
- Broaden "circle of impact" or usefulness
  - Ability to handle any kind of speech

Learning from Human – The Ear?



Learning from Human – The Cortex?

Representation in High Dimensional Space



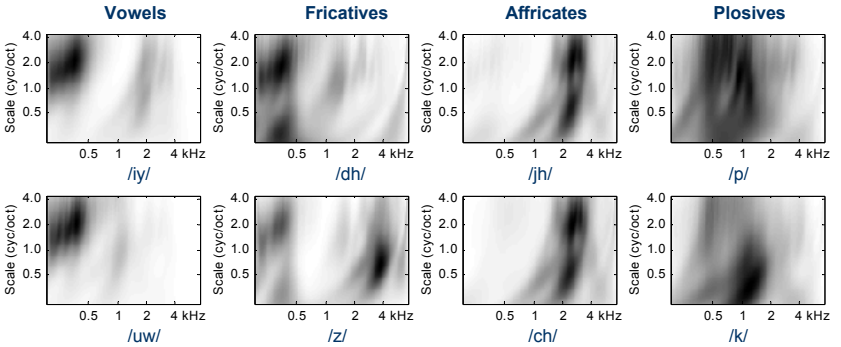
Each response area is parameterized by three variables: x (Best Frequency), s (Bandwidth), ϕ (Symmetry)

Response unit :

$$w_s(y;x,s,\phi) = h_s(y-x)\cos\phi - \hat{h}_s(y-x)\sin\phi$$

Georgia Institute of Technology

"Cortical" Regions of Interest (ROI)



- Regions in the cortical response with **low variance** (light areas above) are likely to contain the identifying features of each phoneme.
- **Class-dependent feature selection and detection**.
- Place coding?

Georgia Institute of Technology

Speech Recognition – Value Proposition

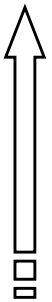
Converse in speech

Understand utterance

"Understand" keywords or limited intents embedded in spoken utterance

Convert speech into word(s)

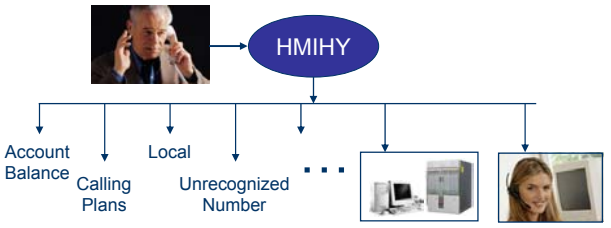
Increasing Value to Users



Georgia Institute of Technology

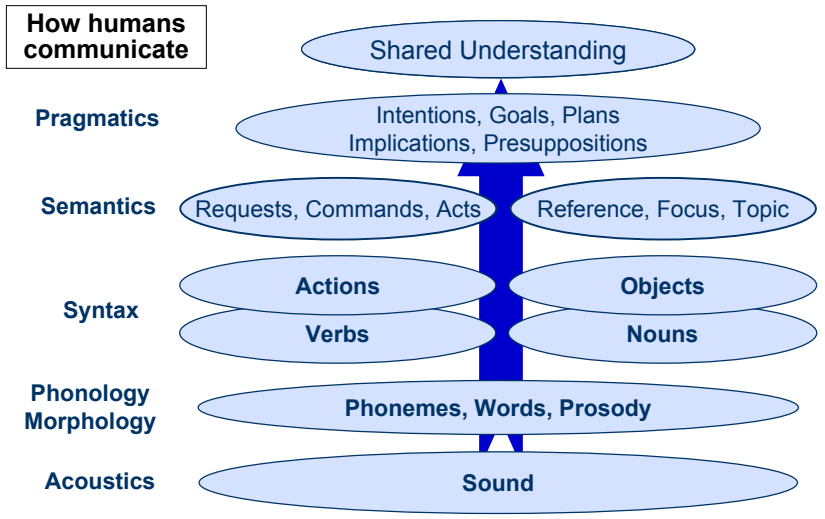
HMIHY & Call Center Routing

- *"AT&T. **How may I help** you?"*
- User responds with ***unconstrained*** fluent speech
  "***`I'd like to know the cheapest calling plan`***"
- System determines the **meaning or intent** in user's speech, then routes the call
- **Dialog** technology enables task completion, cutting 5/6 of transaction time compared to conventional IVR systems



Georgia Institute of Technology

Towards Understanding & Conversation



A Tour for Demo on Use of Reference & Context

Login

Self Awareness I
- What are you doing?

Calendar Arithmetic
- What's the date? … tomorrow's?
- How many days next month?
- What day is the 1st of next month?

Self Awareness II
- What have you done this session?
- What dates have we discussed?

Simple Sequence
- Create <tomorrow eve; 9:00; 30min>
- <Repeat> What are you doing?
- Stop doing that.
- Create a 2-hr appt tomorrow @ noon

Self Awareness III
- What did you just do?
- What day are we talking about?

Selective Access
- What time does that appt end?

Selective Modification
- Extend the duration by one hour.
- Change it back to what it was.
- Resch'ed it for 1st Sat in April.
- Resch'ed it for next Monday.

Conclusion
- Create a 2-hour appt at 2:00 on the last Tuesday of the month after next.
- Goodbye, Daisy.

Speech Everywhere

- Speech is being uttered everywhere, most likely mixed with noise, "competing voice" & distortion
  - **Speech enhancement, noise suppression, de-reverberation?**
- Many sounds exist at any given time – how do we make sense from them?
  - **Source separation? Information fusion?**
- How does the machine know humans are trying to communicate with it?
  - **Modeling of speech act?**
- "Living room" at NTT Communication Science Research

Summary

- Speech technologies are of human communication and for human communication; naturalness is important.
- Speech research has expanded **beyond simple conversion between sound and text**, to wit,

  From Machines That Convert to Machines That Converse
- Speech is an essential transport of intelligence and knowledge; an intelligent machine needs to know how to talk and listen.