Speech Technology: Making computers work naturally with human speech

Alan W Black
Language Technologies Institute
Carnegie Mellon University

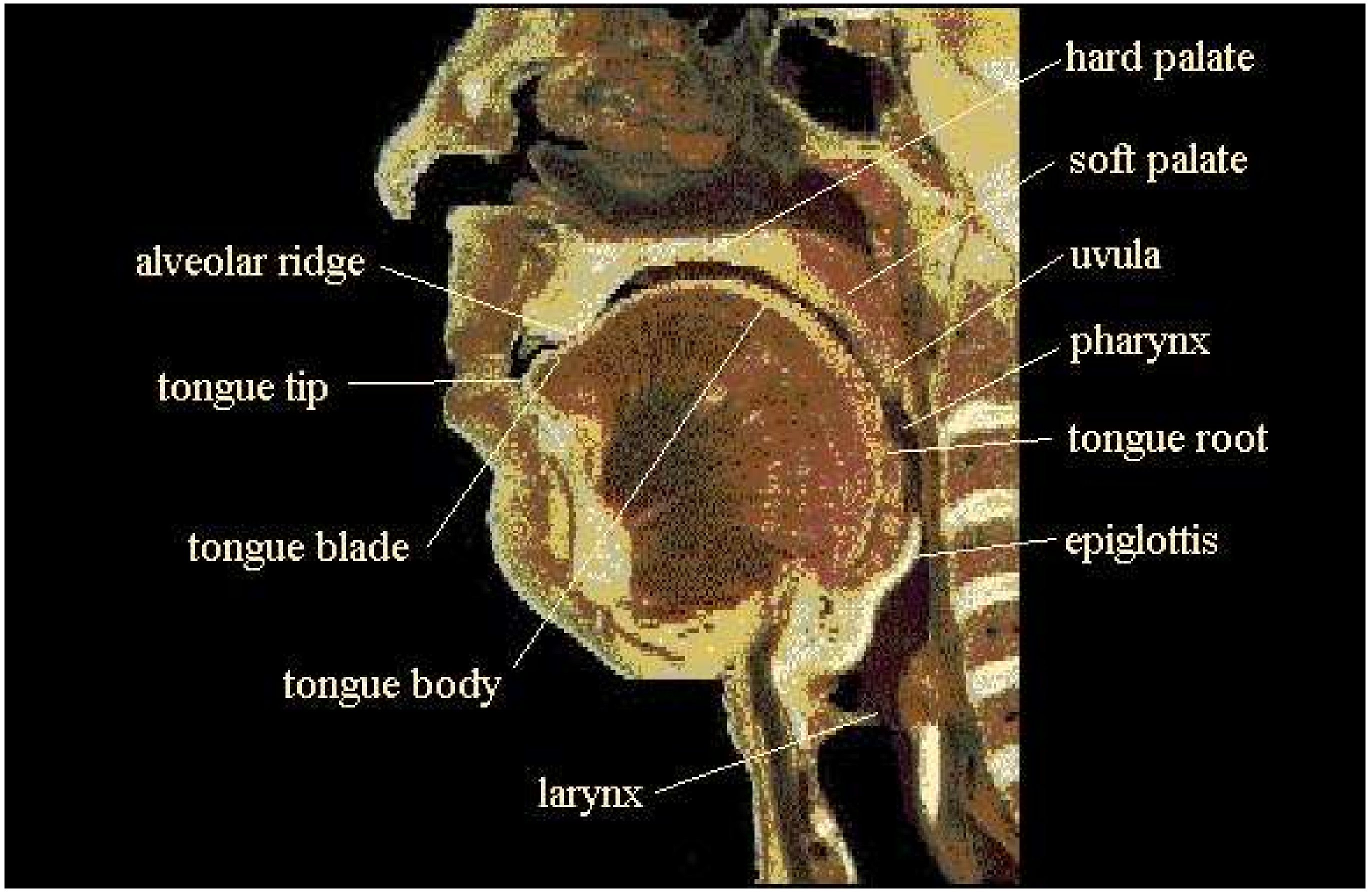
Speech: most natural form of communication

- ☐ Everyone can talk
  – but people have to learn to read and write
- ☐ We can engage in dialog with people through speech:
  – why can't you do that to computers.

But

- ☐ its not good for everything
- ☐ for large amounts of information slow and bulky
- ☐ can't be searched easily
- ☐ its not digital

The vocal tract

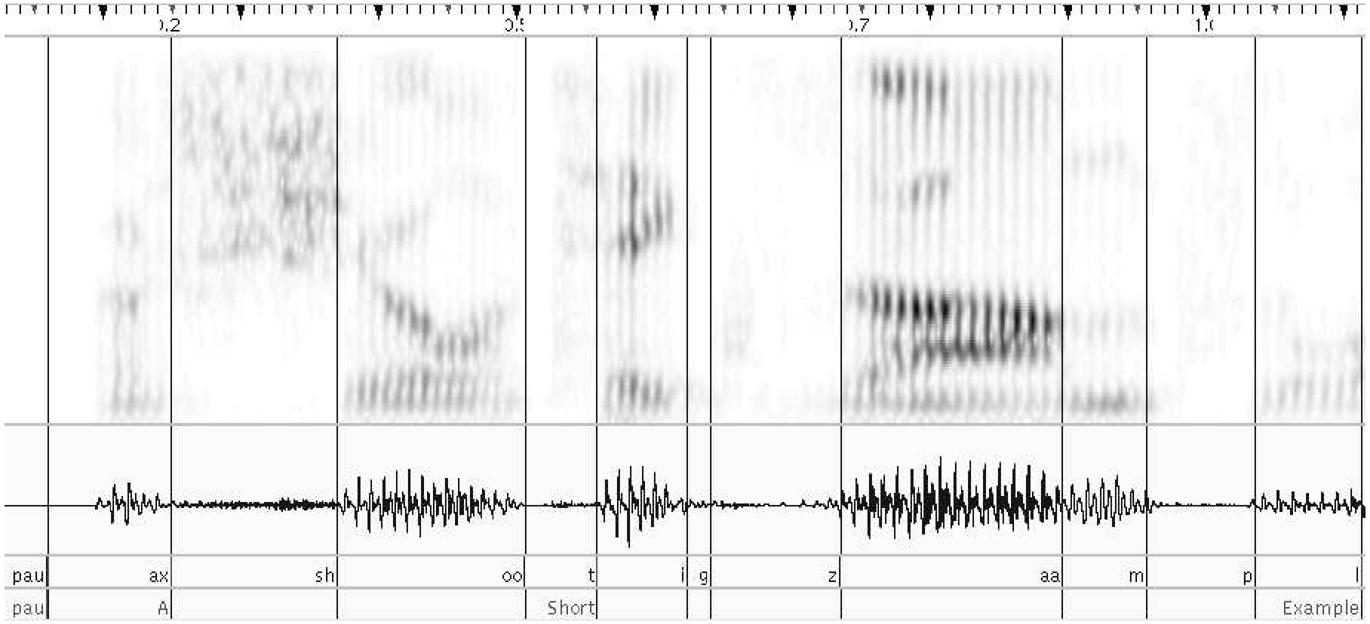


From meat to voice

- From ideas to sound waves:
  - voicing from glottal excitation
  - changing shape of vocal tract
  - obstruents: puting things in the way
  - causes various sound waves to be created
- From sound waves to ideas:
  - sound waves hit your ear
  - flex various hairs in your inner ear
  - brain detects various frequencies
  - magically decodes them

(Note: this trivializes the *understanding* part)

Spectragrams



Linguistics: making it more manageable

- Definition of words:
  - small useful sized objects
- Definition of phonemes:
  - small inventory of sound units
  - different for languages/speaker
- Definition of prosody:
  - phrasing, intonation, durations

Phonology

"smallest unit that when changed (can) change meaning of word."

- "bat" $\rightarrow$ "pat"
- "pat" $\rightarrow$ "pam"

Number of phonemes in a language

- ☐ US English: 43
- ☐ UK English: 44
- ☐ Japanese: 25
- ☐ Hindi: 81

But numbers are not definite

But not all variation is phonological

- ☐ Phonology: linguistic space of sounds:
  – may be a collection of actual sounds
- ☐ Phonetics: "acoustic" space of sounds
  – different sound but not linguistically different

flaps in US English
– "water" → / W AO T ER /
– but common pronunciation / W AO DX ER /

Not all languages are the same

Phonetic variation in one language may be phonological in another

- Asperated stops (Korean, Hindi) P vs PH
- L-R in Japanese not phonological
- US English dialects:
  - mary, merry, marry
- Scottish English vs US English:
  - No distinction between "pull" and "pool"
  - Distinction between: "for" and "four"

Channel Conditions

Different factors affect voice quality

- microphone:
  - head mounted, far field, telephone
- channel:
  - 16KHz/16bit wide band
  - 8KHz/8-12bit telephone
  - 4.8KHz CELP, cell phone
- acoustic conditions:
  - quiet recording studio vs quiet office
  - standing waiting for the bus on a cell phone
  - on an aircraft carrier
- speaker type:
  - regular user
  - new user
  - child/eldery/stressed
  - "value" of information

The key speech technologies

- Speech recognition:
  – taking digital waveforms and producing text
- Speech synthesis:
  – taking text and producing waveforms
- Dialog systems:
  – making this flow in the expected way

Speech Recognition

- ☐ Acoustic parameterization:
  – representing speech invariant of environment
  – time slicing and spectral processing
- ☐ Acoustic modeling:
  – what are all the ways you say "s"
  – HMM modeling
- ☐ Language modeling:
  – what are the most likely words to say
  - "Carnegie ...", "President ..."

Requires "typical" speech to train from

Language modeling: listeners expectations

- In a talk about speech technology:
  - "How to recognize speech with the new display"
- In a news item about a Hawaiian beach:
  - "How to wreck a nice beach with the nudist play"

Markov Modeling

n-gram models:
$P(X_{t+1} \mid X_t, X_{t-1}, ...X_{t-n})$

- From data collect all n-gram distributions
- Need lots of data
- Need to smooth it:
  - “green table” never appeared in WSJ1995

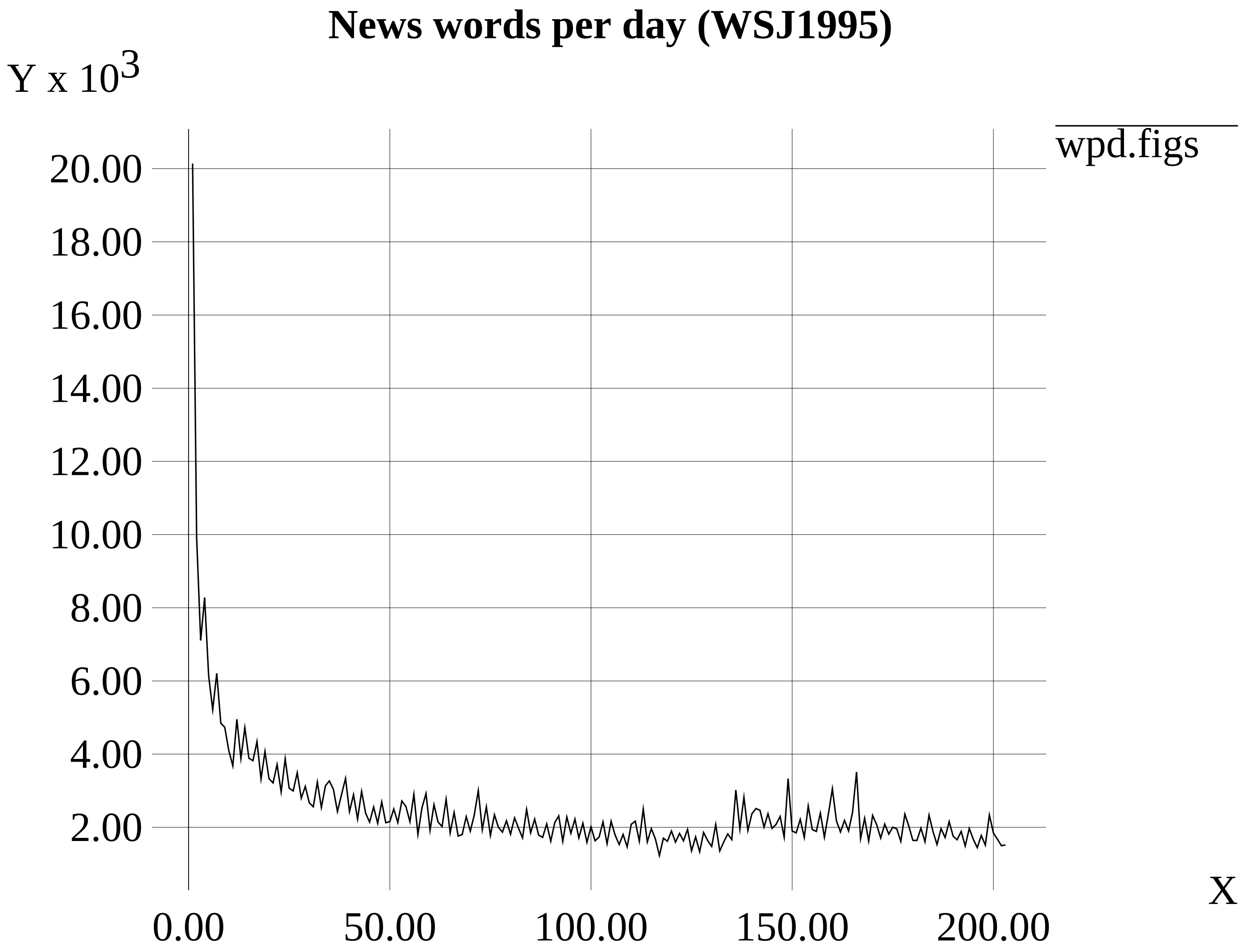
One year of WSJ (1995)

Total of 22.5M word tokens

Total of 508K word types

15K types appear more than 100 times

45% tokens appear only once


News words per day (WSJ1995)
Y x 10^3
wpd.figs
20.00
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
50.00
100.00
150.00
200.00
X

Using LM in ASR

- ☐ find word sequence that maximizes $P(W \mid O)$
- ☐ Using Bayes' Law

$$\frac{P(W)P(O|W)}{P(O)}$$

- ☐ Combine models:
  – Use HMMs to provide $P(O \mid W)$
  – Use language model to provide $P(W)$
- ☐ Some grammar factor weight

Speech Synthesis

- Find out what to say:
  – get pronunciations of words, token etc
- Add prosody:
  – make it not be a boring monotone
- Make a waveform by:
  – concatenating small pieces of pre-recorded speech

Homographs

Words with same written form but different pronunciation

- ☐ Different part of speech: project
- ☐ Semantic difference: bass, tear
- ☐ Proper names: Nice, Begin, Said
- ☐ Roman Numerals: Chapter II, James II
- ☐ Numbers: years, days, quantifiers, phone numbers
- ☐ Some symbols: 5-3, high/low, usr/local

How common are they?
– Numbers: email 2.57% novels 0.00013%
– POS/hgs: WSJ 7.6%

Homograph disambiguation (Yarowsky)

Same tokens with different pronunciation

- □ Identify particular class of homographs
  – e.g. numbers, roman numerals, “St”.
- □ Find instances in large db with context
- □ Train decision mechanism to find most distinguished feature

Homograph disambiguation: example

Roman numerals: as cardinals, ordinals, letter

Henry V: Part I Act II Scene XI: Mr X is I believe, V I Lenin, and not Charles I.

- Extract examples with context features
- Label examples with correct class:
  – king, number, letter
- Build decision tree (CART) to predict class

Features

```
class: n(umber) l(etter) c(entury) t(imes)
rex rex_names section_name num_digits p.num_digits n.num_digits
pp.cap p.cap n.cap nn.cap



n II 0 0 0 11 7 2 3 7 0 0 1 1
n III 0 0 0 3 4 3 3 5 0 0 1 1
c VII 1 0 0 4 9 3 3 3 1 1 0 0
n V 0 0 1 3 4 1 1 2 0 1 0 1
n VII 0 0 1 2 4 3 1 2 0 1 0 1
...
```

```
      ((p.lisp_tok_rex_names is 0)
       ((lisp_num_digits is 5)
        ((number))
        ((lisp_num_digits is 4)
         ((number))
         ((nn.lisp_num_digits is 13)
...
                   ((nn.lisp_num_digits is 2)
                    ((letter))
                    ((n.cap is 0) ((letter)) ((number))))))))
...
```

Homograph disambiguation: example

Example data features:
– surrounding words, capitalization, “king-like”, “section-like”

class	ord	let	card	times	total	correct	percent
ord	133	0	15	0	148	133/148	89.865
let	3	40	9	0	52	40/52	76.923
card	7	6	533	0	546	533/546	97.619
times	0	2	1	1	4	1/4	25.000

707/750 94.267% correct

Homograph disambiguation

But it still fails on many obscure (?) cases

- ☐ William B. Gates III.
- ☐ Meet Joe Black II.
- ☐ The madness of King George III.
- ☐ He's a nice chap. I met him last year.

How many homographs are there?

Very few actually, ...

axes bass Begin bathing bathed bow Celtic close cretan Dr executor jan jean lead live lives Nice No Reading row St Said sat sewer sun tear us wed wind windier windiest windy winds winding windily wound Number num/num num-num Roman_numerals

Plus *many* POS homographs

Unit selection synthesis

- Select appropriate units of speech from database of natural speech
- What are the unit size:
  – halfphones, phones, syllables, words
- How large should it be:
  – design a database with coverage
- How to select them
  – what distance functions to use

Clustering for Unit Selection

Black and Taylor, Eurospeech '97

Find mean acoustic distance between all units of the same class (e.g. phoneme type)

$$Impurity(C) = \frac{1}{|C|^2} * \sum_{i=1}^{|C|} \sum_{j=1}^{|C|} Adist(C_i, C_j)$$

Recursively find best question which splits C (cf. CART) so mean impurity of sub-clusters less than impurity of C.

Questions used:
- phonetic context
- pitch and duration context
- syl position, stress, accent
- position in phrase

Acoustic measure:
- Mel cepstral coefficients
- F_0 and power

Runtime synthesis selects appropriate cluster with CART tree and find best path through candidates with Viterbi search.

Dialog systems

- Who's turn is it
- What the current topic:
  – what does "it" refer to
- Is the dialog directed:
  – is there a goal, are we getting to it
- What is the state:
  – was a question asked/answered
  – was the phrase relevant

What are the key uses

- ☐ Command and control
- ☐ Spoken dialog systems:
  - (telephone-based) information services
- ☐ Information retrieval from audio:
  - tell me all CNN broadcasts about WorldCom
  - meeting summarization
- ☐ Speech-to-Speech translation:
  - device that will translate
- ☐ Computer aided education:
  - language training
- ☐ Interactive agents:
  - robot characters that talk with you

Making the computer talk in your voice

`http://festvox.org/`

- Tools, documentation, aligners, and scripts
  - Build your own voice synthesizer
  - US and UK English diphone synthesizer (1-2 days)
  - Other languages (1 week to ... much longer)
- Building a voice:
  - record *appropriate* speech in *appropriate* style
  - build unit selection synthesizer
- Different techniques:
  - recorded prompts
  - limited domains
  - general voices
- In English or other languages

Speech Synthesis Components

- I want my computer to talk
  - Speech Synthesis Engine
  - Festival Speech Synthesis Systems
  - converts text to speech in English and other languages
- I want my computer to talk in my voice
  - tools for building new voices
  - The FestVox project
  - general and domain voices
- I want my voice on my PDA/Cell phone now
  - Small footprint synthesis
  - CMU Flite
  - Client based content delivery systems

Make it sound better

- General voices
  - Say anything
  - word concatenation
  - phone concatenation
  - diphone concatenation
  - unit selection synthesis
- Domain voices:
  - targeted to a domain
  - much higher quality:
  - clocks, weather, stocks, simple dialogs

Make it smaller and faster

- General voices
  - large requiring big servers
  - greater than 1GB memory
- Small footprint synthesis:
  - small memory, processor requirements
  - no compromise on quality

USI: Universal Speech Interface

`http://www.cs.cmu.edu/~usi/`

A common, easy-to-learn interface to speech applications

- Choice:
  - make you speech interface accept anything, or
  - spend a little time to educate you user to a standard
- Like “Graffitti” for Palm:
  - not standard writting
  - but easy to learn
  - and easy to recognize
- `http://www.speech.cs.cmu.edu/usi`

Communicator: mixed initiative spoken dialog

http://www.speech.cs.cmu.edu/Communicator

- □ DARPA funded project with multiple site:
  – MIT, Colorado, AT&T, Lucent etc
- □ Telephone based access to flight information :
  – call 412 268 1084 (1-877-CMU-PLAN)
- □ Any speaker
- □ Mixed-initiative
- □ Accessing live data on the web

CSTAR: speech to speech translation

`http://www.c-star.org/`

Joint effort with 16 other sites worldwide

- Speech translation in the tourism information domain
- "Can you tell me the way to the conference center"
  - Kaigi sentaa no hou ga oshiete kudasaimasen ga
- Includes:
  - English, German, Italian, Korean, Japanese, ...

DARPA Babylon project

- ☐ Hand held, portable speech-to-speech translation
- ☐ "One way"
  - – fixed phrase translation
  - – answers can be yes, no and pointing
- ☐ "One+One way"
  - – fixed phrase translation both ways
- ☐ Two way:
  - – constrained but general speech
  - – Medical triage, Refugee Processing, Force protection

In languages with little cuurernt support:
– Pashto, Dari, Farsi and Arabic.

Meeting summarization

Record a meeting and annotate it with who said what

- ☐ More than one speaker at once
- ☐ People may move, arrive, leave
- ☐ Voices may get heated
- ☐ Audio "grep":
  - "find bits where Fred complained about Q1 figures"

Some difficult speech problems

- ☐ How do you deal with real speech input
- ☐ How do you teach the users what they can say
- ☐ How do you present to the user complex information
- ☐ How can you make it fast enough
- ☐ How do you mix speech and graphics
- ☐ How do you make dialogs work in new domains/languages

Future speech applications

- Singing synthesis:
  - would you like to sing along to ...
- Interactive agents:
  - Personal Digitized Assistants
  - information gatherers and presenters
- Speech based question and answering:
  - auto-FAQ by telephone
- Speech will become default interaction language