Speech Technology

Making computers work naturally with speech

Alan W Black
Language Technologies Institute
Carnegie Mellon University

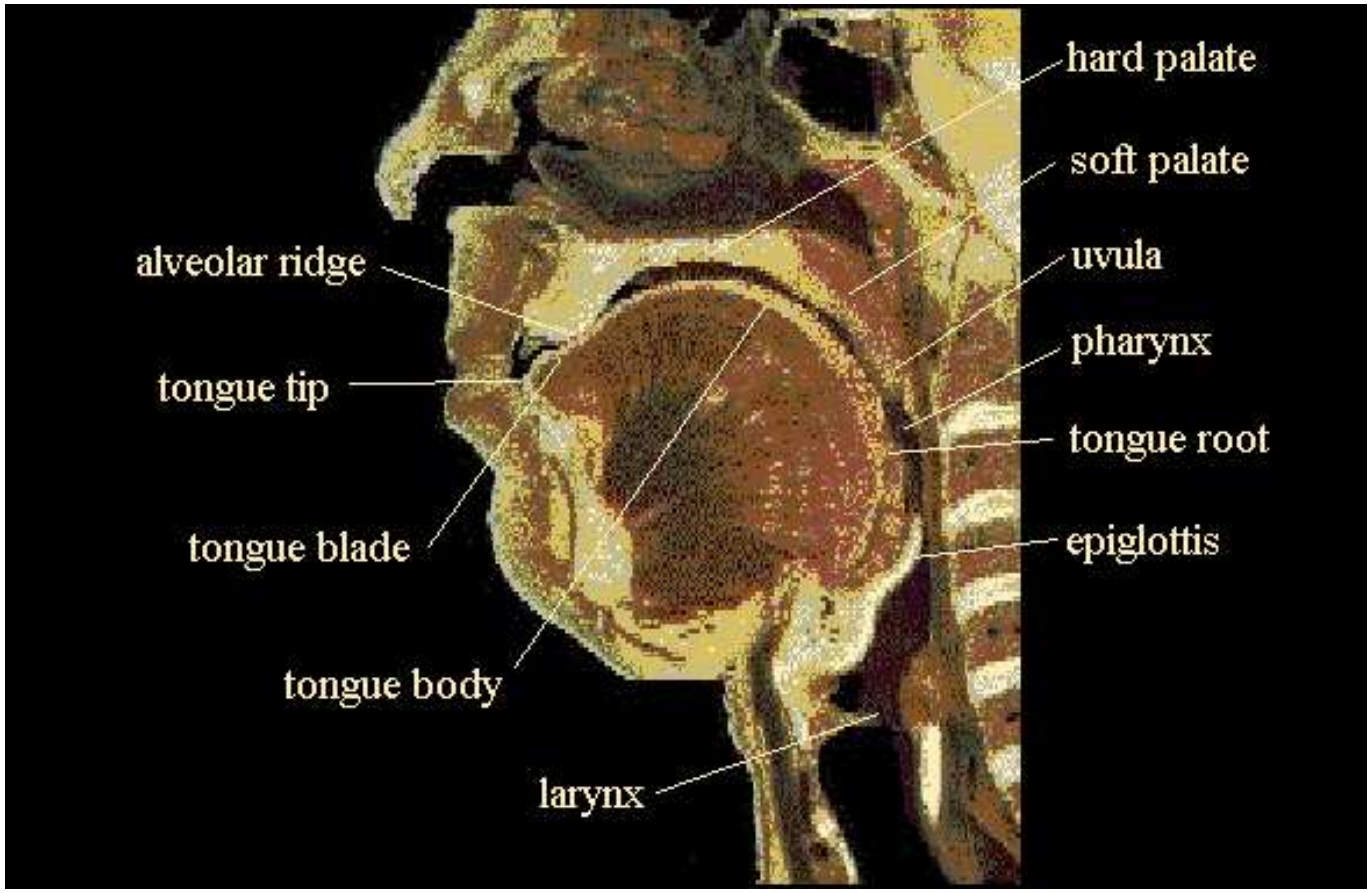
Speech: most natural form of communication

- 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

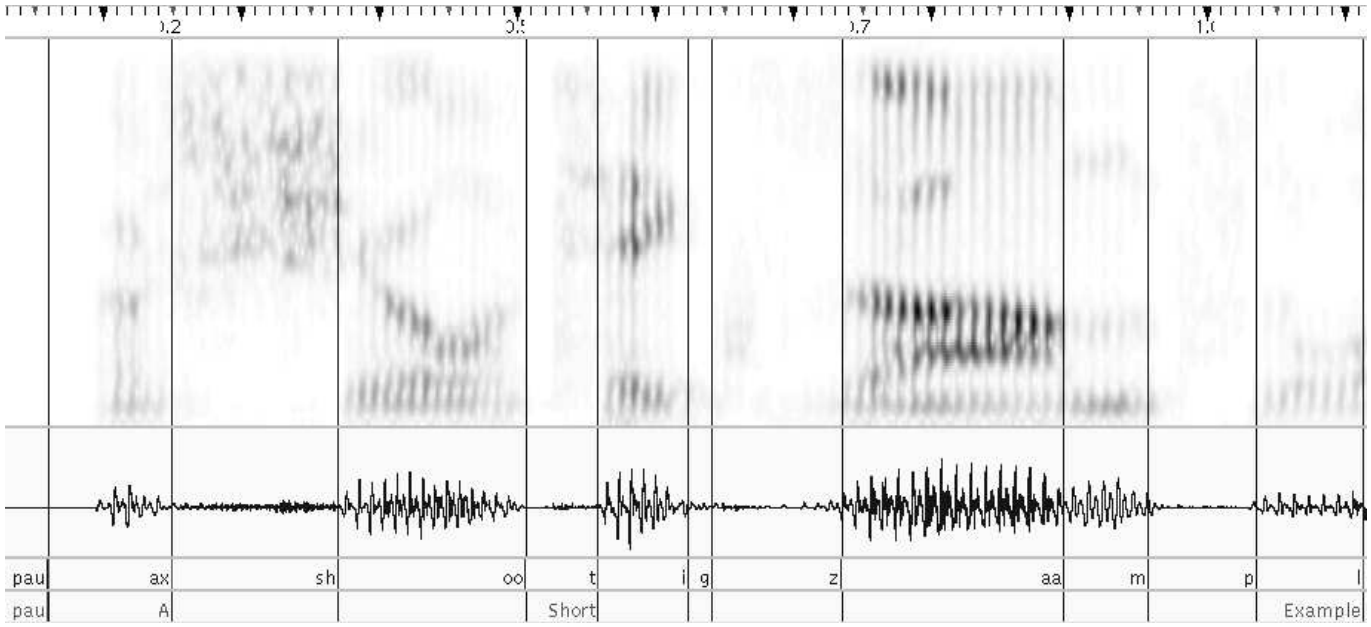


From meat to voice

- 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"

US English Vowels

AA	wAshington	AE	fAt, bAd
AH	bUt, hUsh	AO	lAWn, mAll
AW	hOW, sOUth	AX	About, cAnoe
AY	hIde, bIble	EH	gEt, fEAther
ER	makER, sEARch	EY	gAte, Ate
IH	bIt, shIp	IY	bEAt, shEEp
OW	lOne, nOse	OY	tOY, OYster
UH	fUll, wOOd	UW	fOOl, wOOd

US English Consonants

- Stops: P,B,T,D,K,G
- Fricatives: F,V,HH,S,Z,SH,ZH,TH,DH
- Affricatives: CH, JH
- Nasals: N, M, NG
- Glides: L, R, Y, W

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" $\rightarrow$ / 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 modelling:
  - what are all the ways you say "s"
  - HMM modelling
- Language modelling:
  - what are the most likely words to say
  - "Carnegie ...", "President ..."

Requires "typical" speech to train from

Language modelling: listeners expectations

Last Saturday in Hawaii, numerous Waipouli vacationers were shocked to find their beach cordoned off for a UC Berkeley Drama enactment of "Personal office space". The play features exclusively topless men and women in an everyday office environment. Richard Carlson, one of the annoyed tourists and a regular swimmer at Waipouli beach, complained that they really knew how to wreck a nice beach with the nudist play. Many of the tourists appeared ruffled by the content and fled the scene to avoid compromising photos.

Language modelling: listeners expectations

In yesterday's press release, AT&T unveiled SpeechKit, its new speech recognition toolkit. According to Michael Armstrong, the COO of the company, the most innovative feature of the system is its revolutionary three-dimensional interface, which opens a new universe of possibilities for the speech recognition community. During the official software release, Jonathan Blues, a senior researcher at AT&T Labs, explained how to recognize speech with the new display, and how the toolkit has already played a crucial role in his research.

Language modelling: listeners expectations

- how to recognize speech with the new display
- how to wreck a nice beach with the nudist play

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

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

Fluency: computer aided language learning

`http://www.lti.cs.cmu.edu/Research/Fluency/`

- Coaching for "difficult" phonemes
  - "th", vowel length
- Have language learners speak utterances
- compare against "golden voice"
- find duration, F0, stress, spectral problems
- Hard problem of recognizing non-native speech

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"

Project Listen: teaching children to read

`http://www.cs.cmu.edu/ listen/`

- Use speech recognition to listen to children reading:
  - detect errors
  - read passages to students
- Follow progress and improve reading standard
- Attacking hard problems:
  - how do you recognize non-fluent children's speech
  - how do you know if the system works

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