Current issues in Speech Technology

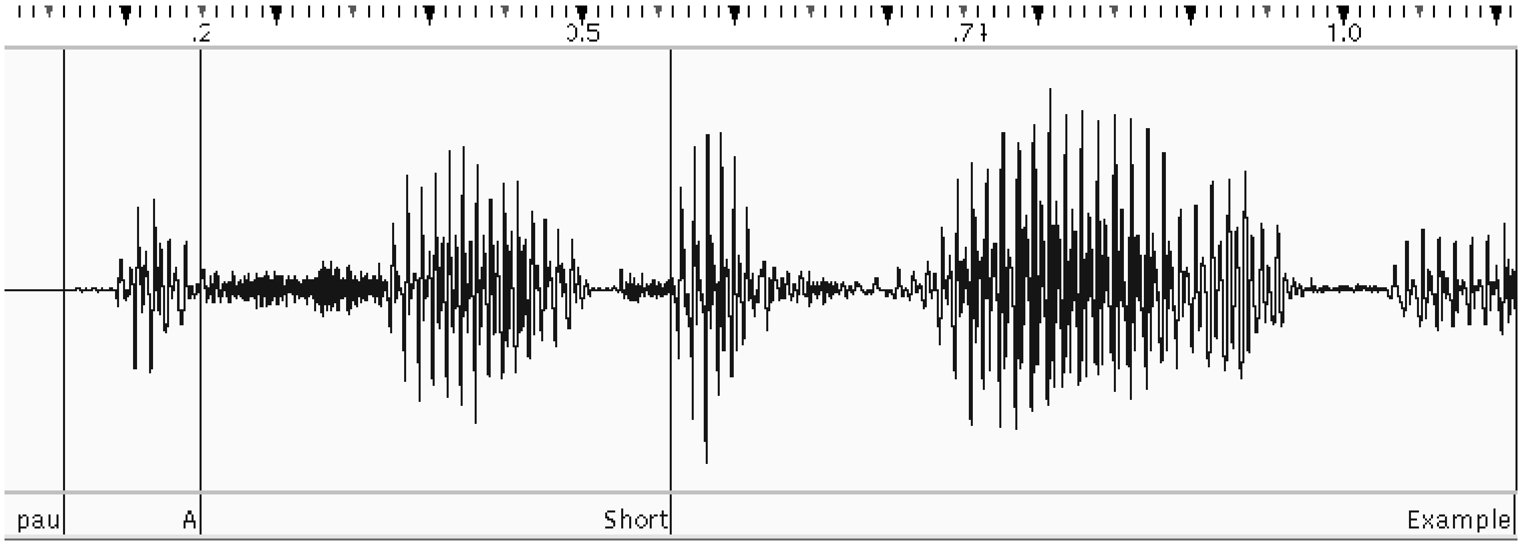
Alan W Black
Language Technologies Institute
Carnegie Mellon University

Overview

- ☐ Computers and speech
- ☐ Speech Recognition
  - Acoustic parameterisation
  - Language modelling
  - How well does it work
- ☐ Speech Synthesis
  - Text preocessing, pronucniation, prosody
  - Waveform generation
- ☐ Some applications
  - Speech to Speech translation
  - Talking heads
  - Dialog Systems

Digitized speech

- ☐ Sample A/D converter at set freqency
  - 8Khz: telephone speech
  - 16Khz: high quality voice
  - 44.1Khz: CD quality



Speech components

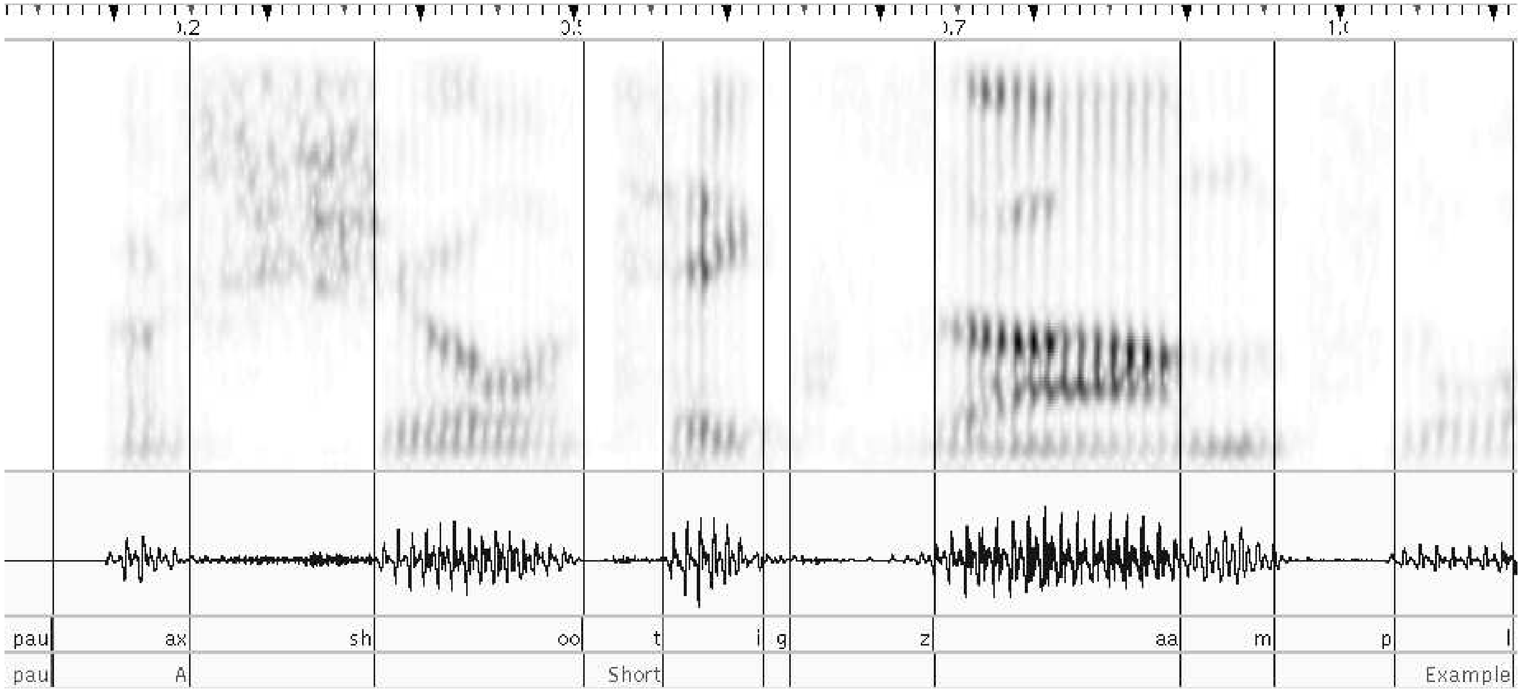
- Phonemes: linguistically distinct units
  - *desk* → d eh s k
  - *car* → k aa r
  - *care* → k eh r

- Phonetic classes:
  - voiced/unvoiced
  - nasals: m n ng
  - stops: t d k g b t
  - fricative: s z sh zh f v
  - etc

Phones as frequency distrbutions

- ☐ Looking at *amplitude* and *duration* is limited
- ☐ Look at frequency components
  - – sh $\rightarrow$ "white noise", multi-frequency
  - – ii $\rightarrow$ clear frequencies (formants)
- ☐ Can convert amplitude (time domain) into frequency doamin
  - – `fast fourier transform` (FFT)
  - – Each sample made up of combination of sine waves

Viewing speech in frequency domain



Speech recognition

- ☐ How to wreck a nice beach.
- ☐ How to recognize speech.

Speech recognition

- ☐ Template match in time domain:
  - – small vocabulary
  - – speaker specific

- ☐ Constrained phone recongition in freq. domain:
  - – larger vocabulary
  - – speaker independent

Types of speech recognition

- Isolated word
  - yes/no, digits
  - command and control
- Word spotting:
  - "... computer, locate Commander Riker ..."
  - AT&T call direction
- Connected speech:
  - pauses between words
- Continuous speech:
  - dictation (Dragon Dictate)
  - transcription
  - dialog

Mordern speech recognition architecture

- Phone(me) recognizers:
  - *hidden markov models* (HMM)
  - trained statistical model
  - probability of input being some phone
- Language model:
  - most likely word sequence
  - hence most likely phone sequence

Source channel model

- Find best word sequence given acoustic observations:

$$\mathbf{w} = \operatorname{argmax} p(\mathbf{w}|\mathbf{o}) \tag{1}$$
$$= \operatorname{argmax} p(\mathbf{o}|\mathbf{w})p(\mathbf{w}) \tag{2}$$

- $\mathbf{o}$ is HMM output from FFT frames
- $\mathbf{w}$ is $p(w_i|w_{i-1}, w_{i-2})$

Language modelling

- ☐ What word comes next ...:
  - *Carnegie* ...
  - *Water* ...

- ☐ Estimate probability of next word
  - $p(w_i|w_{i-1}, w_{i-2}, ...)$
  - Collect **large** databases
  - Find statistics of word frequencies
- ☐ What about new words
- ☐ What about larger context

Current systems

- ☐ CMU's Sphinx:
  - large vocabulary speaker independent
  - multiple languages
- ☐ Dragon, IBM, AT&T, Nuance, Cambridge
- ☐ Realtime and Offline:
  - 5,000-150,000 word vocabulary.

Current tasks

- Dictation:
  – IBM's ViaVoice, Dragon Dictate
- DARPA Communicator
  – Telephone dialog systems
- Broadcast News
  – Radio/TV transcription
- Speech-to-Speech Translation
  – CStar project (ATR, CMU, ...)
  – Verbmobil
- Voice portals
  – news, sports, email etc – e.g. quack.com, tellme.com etc

Speech Synthesis

- Making computers talk.
- Converting text to speech:
  - email messages, news reports
- Converting "concept" to speech:
  - Structured text to speech
  - Output from natural language generation systems

Speech Synthesis Tasks

- Finding the words in text
- Giving the pronunciation of words
- Generate **prosody**:
  – phrasing, intonation, duration and power
- Generate waveforms:

Text analysis

- This is a pen.
- My cat who lives dangerously has nine lives.
- He stole $100 from the bank.
- He stole 1996 cattle on 25 Nov 1996.
- He stole $100 million from the bank.
- It's 13 St. Andrew St. near the bank.
- Its a PII 450Mhz, 128MB RAM, 4.5Gb SCSI, 24x cdrom and 19" monitor.
- My home pgae is `http://www.geocities.com/awb/`.

Text analysis - examples

- Prosodic phrasing using part of speech
  - *He wanted to go for a drive in*
  - *He wanted to go for a drive in the country*
- Homograph disambiguation
  - *My cat who lives dangerously had nine lives.*
  - *Henry V: Part I Act II Scene XI: Mr X is I believe, V I Lenin, and not Charles I.*

```
<?xml version="1.0"?>
<!DOCTYPE SABLE PUBLIC "-//SABLE//DTD SABLE speech mark up//EN"
        "Sable.v0_2.dtd"
[]> <SABLE> <SPEAKER NAME="male1">

The boy saw the girl in the park <BREAK/> with the telescope.
The boy saw the girl <BREAK/> in the park with the telescope.

Some English first and then some Spanish.
<LANGUAGE ID="SPANISH">Hola amigos.</LANGUAGE>
<LANGUAGE ID="NEPALI">Namaste</LANGUAGE>

Good morning <BREAK /> My name is Stuart, which is spelled
<RATE SPEED="-40%">
<SAYAS MODE="literal">stuart</SAYAS> </RATE>
though some people pronounce it
<PRON SUB="stoo art">stuart</PRON>.  My telephone number
is <SAYAS MODE="literal">2787</SAYAS>.

I used to work in <PRON SUB="Buckloo">Buccleuch</PRON> Place,
but no one can pronounce that.

By the way, my telephone number is actually
<AUDIO SRC="http://att.com/sounds/touchtone.2.au"/>
<AUDIO SRC="http://att.com/sounds/touchtone.7.au"/>
<AUDIO SRC="http://att.com/sounds/touchtone.8.au"/>
<AUDIO SRC="http://att.com/sounds/touchtone.7.au"/>.
```

SABLE: for marking emphasis

What will the weather be like today in Boston?

`It will be <emph>rainy</emph> today in Boston.`

When will it rain in Boston?

`It will be rainy <emph>today</emph> in Boston.`

Where will it rain today?

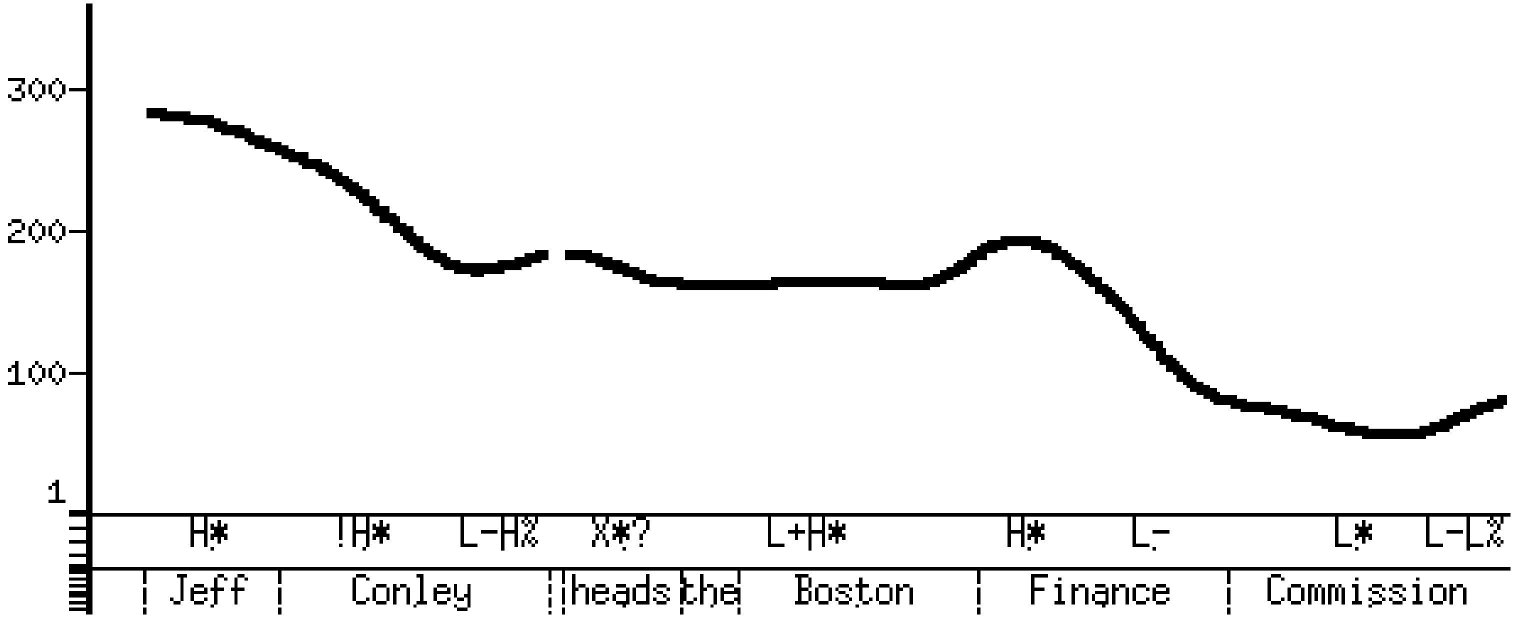
`It will be rainy today in <emph>Boston</emph>.`

Prosody

- ☐ Phrasing
  - chunking in speech
- ☐ Intonation
  - Accents and F0 generation
- ☐ Duration:
  - length of segments
- ☐ Power:
  - loudness

Each has underlying "default", and marked forms

Intonation marking



Intonation Examples

Getting there without training data

- ☐ Fixed durations, flat F0.
- ☐ Decline F0
- ☐ "hat" accents on stress syllables
- ☐ accents and end tones
- ☐ statistically trained

Waveform generation

- ☐ Articulary synthesis:
  - – model the vocal tract
- ☐ Formant synthesis:
  - – construct the signal from articial components
- ☐ Concatenative synthesis:
  - – record human speech
  - – piece together appropriate parts

Waveform generation

- ☐ Random word/phrase concatenation
- ☐ Phone concatenation
- ☐ Diphone concatenation
- ☐ Sub-word unit selection
- ☐ Cluster based unit selection

Prosodic modification of speech

- Independently modify pitch and duration
- Modifiy pitch:
  - place pitch periods closer together/further apart
- Modifiy duration:
  - duplicate or delete segments

– slow/fast
– high/low
– flat/wide

Diphone synthesis

- All phone-phone transitions in language
- Mid-phone is more stable so easier to join
- Record all diphones in nonsense words:
  For example
  `kal_0001 ("b-aa" "aa-b") (t aa b aa b aa)`
  `kal_0002 ("p-aa" "aa-p") (t aa p aa p aa)`
  1348 words for US English (1618 diphones)

Unit selection synthesis

- Large representative database
- Select "appropriate" units:
  - phonetically and prosodically
- Various unit selection algorithms:
  - target, concatenation cost
  - cluster
  - phonological structure matching

Cluster method

Unittypes: phones
– acoustically partitioned.

Occurrences: many.

Target distance: acoustic distance from cluster centre.

Concatenation distance: dynamically determined
(optimal coupling)

Unit size: uniform

Selection features: phonetic, prosodic, positional

Building voices automatically

- ☐ Define, record and label diphone set
- ☐ Build lexicon, letter to sound systems
- ☐ Build prosodic models from data or rules
- ☐ Build text analyser

Current languges (in Festival):
– English (US and UK), Spanish, German, Welsh
– Greek, Polish, Basque, Korean, Swedish, Klingon

Dialect synthesis

- ☐ Different dialects have different styles:
  – phones, prosody, vocab and grammar
- ☐ Build lexicon with "key vowels":
  – the vowel in *coffee*
- ☐ Record person's "coffee" vowel ...

Dialect synthesis:
– RAB, british RP English
– JGM, Southern Irish English
– KED, Standard US English

– I say Pakistan (RAB,JGM)
– What a waste (RAB,JGM)
– Its hotter in the city (JGM,KED)

DARPA Communicator System

- ☐ Telephone-based dialog systems
- ☐ Airline, hotel and car registration
- ☐ Structured but free dialog
- ☐ Access web for *real* information

Telephone Speech Applications

- CMU Airline reservation: 268-1084 (268-5144)
- CMU Movieline: 268-1185
- MIT Weather 1-888-573-8255
- quack.com 1-800-737-8225

Other synthesis uses

- Dialog systems
- Talking heads (OGI/Fluent Tech/UCSC)
- Singing (OGI/Fluent Tech)

End comments

- Speech technology almost works
- Some people find speech *cool*
  – but others find it not helpful
- If it works well enough
  – people will use it.