Beyond Text to Speech

- un-restricted text
- style dependent analysis
  - e.g email, address list, etc
  - special tokenizing
  - lexicon, prosody
- explicit mark-up
  - identifying phrases
  - basic structure
- structure input
  - Concept to speech
  - syntactic/semantic/pragmatic information
- Other
  - limited domain
  - canned waveforms
  - phonetic specification
  - intonation specification

Style specific processing

- email:
  - header analysis
  - quote processing
  - more abbrevs
- addresses:
  - slower, more prosodic breaks
  - number analysis
  - special abbrevs

Reading addresses

Smith, Bobbie Q, 3337 St Laurence St, Fort Worth, TX 71611-5484, (817)839-3689
Anderson, W, 445 Sycamore Way NE, Lincoln, NE 98125-5108, (212)404-9988

```
<?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>.`

But we need a richer markup

- SABLE is quite limited
- Concept to speech is richer:
  - translation and generation systems
  - Syntactic, Semantic
  - Anaphoric, Rhetorical, Speech act etc.
- Mark up should be:
  - abstract not low-level
  - e.g *type=question* not
  - *pitch rise at end*

SOLE: Spoken Output Intelligent Labeller

- Describe exhibits in museum
- Context and user sensitive
- Practical aspects

Combine:

- ILEX: language generation system
- Festival Speech Synthesis System

Example

This necklace is made of silver, also made by Anne Morris. It is decorated with deep red garnets held by copper and silver alloy wire and a pink cut-glass disc over the clasp. These colours were used to show solidarity with the suffragette movement.

SOLE: intonation prediction

- ☐ Syntactic, Semantic
- ☐ Rhetorical, Anaphoric
- ☐ Also:
  - – proper names, dates etc.

Accent placement prediction:
– tested on held out data
– SOLE info reduced errors by 15%

Speech generation

- ☐ What are the important labelling of speech/language:
  - – dialog act type
  - – emphasis, focus
  - – contrast
- ☐ How do we render this prosodically:
  - – phrasing
  - – accent placement (and non-placement)
  - – contour types etc
- ☐ Lexical choice:
  - – audio channel is different from text/graphics
  - – List of numbers in audio is difficult
- ☐ Identify confusables:
  - – 14 (one four), 25
  - – 20 5 25
  - – "Campbell" with a "p"

Customize: Synthesizer or Text

Two directions

- ☐ Customize the synthesizer:
  - – specific modes
  - – tailor process for the domain
- ☐ Customize the text:
  - – add explicit markup yo text
  - – add explicit processing for markup

Clearer Spoken Output (Let's Go)

- The elderly can't understand synthetic speech:
  - so make it shout
  - problem: familiarity or hearing?
- Lexical Choice:
  - *The next bus is at 10:23*
  - *The next bus is in 11 minutes*
- Prosodic choice:
  - phrasing. duration and intonation
  - *The next bus is at 10:23*
  - *The next bus is at, 10:23*
- Spectral characteristics:
  - clear articulation (not "shouting")
  - *The next bus is at, 10:23*

Synthesis Evaluation

- How good is a voice?
- Is voice X better than voice Y
- Why

Human tests

- Synthesis people are warped:
  - the more you listen to a voice the easier it is to understand
  - they hear things in the voice others don't
- non-synthesis people will be warped:
  - people are very sensitive to experiment conditions
  - what question do you ask
  - hardware you play on makes a difference
- These seem to be orthogonal:
  - understandable
  - natural

Standard Tests

- ☐ DRT: diagnostic rhyme test
  - – tests confusable phones in nonsense words
  - – "bat", "pat", "mat", "kat"
  - – good for identifying phonetic quality
  - – but harder to test in unit selection
- ☐ SUS: semantically unpredictable sentences:
  - – det adj noun verb det adj noun pre det adj noun
  - – "The inedible chair ate a tartan banana with a happy book"
- ☐ Scoring results:
  - – MOS: mean opinion scores
  - – 1-5 quality, naturalness, ...
  - – Take average score

Some experimental problems

- Order of presentation
- Other aids changes perception:
  - showing the text makes it much easier
  - having a talking head improves the synthesis
- Hardware quality:
  - some voices are better on the telephone
  - speaker (hardware) quality
  - room acoustics
  - volume
- Understandability:
  - Difference when listening comfortably vs
  - listening in noisey environment
- Personal preference:
  - voice is fully understandable but "creepy"
  - voice is incomprehensible but "funny"

But how good are your ears?

Can you hear the difference

- “In mud eels are, in clay none are.”
- A 1918 state constitutional amendment,
  made Massachusetts one of twenty-three states,
  where citizens can enact laws by plebiscite.
- Which is which:
  - “The numbers are 25 and 34.”
  - “The numbers are 20 5 and 34.”
- What is the temperature in Pittsburgh?

Objective Synthesis Tests

- Text analysis
  - how well do you cover NSW
  - how well do you cover homographs
- Lexical coverage
  - how often do you see new words
- Lexical correctness
  - how correct are your words
  - for unseen/for "known"
- Phonetic intelligibility
  - DRT type tests
- Semantic intelligibility
  - SUS tests

What about prosody?

Embedded Speech Synthesis

- Speech is large:
  - difficult to transfer over bandwidth limited networks
  - processing locally would be good
  - transfer text and/or phone/dur/F0
- Applications:
  - information services
  - very low bit voice transfer
  - on cell phone and/or pda

How small can you make a synthesizer

- Festival (standard):
  - Binary: 4M
  - Lexicon: 5M
  - Diphones: 8M
  - Runtime memory 20-30M
- Festival (careful):
  - Binary: 2M
  - Lexicon: 3M
  - Diphones: 1.5M
  - Runtime memory 16M
- Flite (Festival-lite):
  - Binary: 50k
  - Lexicon: 0.5M
  - Diphones: 0.75M
  - Runtime memory 1M

Reducing resources

- Lexicon:
  - LTS as FSTs
  - exception list as FSTs
  - minimised
- Diphones:
  - remove similar diphones (e.g. phone_STOP)
  - spike excited LPC
  - quantised LPC frames

Will it be fast enough

- Festival runs at about 0.05 realtime:
  - a 10 sec utterance takes 0.5 sec to synthesize
  - But a 60 second utterances ..
  - OK for reading a book
  - Can spool the audio
- Time to first audio:
  - for dialog must be less than 250ms
  - But data transfer, audio device set up costs too
- Utterance by Utterance:
  - phrase by phrase
  - direct