Research issues in Speech Synthesis

Alan W Black
Language Technologies Institute

Overview

- Some issues in speech synthesis
- The Festival Speech Synthesis System
- Components of a voice
  - Text analysis and lexical lookup
  - Prosodic modeling
  - Waveform synthesis
- Some interesting projects

Aspects of speech synthesis

- Apparently easy
  – get pronunciation, put sounds together.
- Small examples can sound good
- Everyone views the problem differently
  – waveform synthesis (concatenation/formant)
  – prosody
  – pronunciation
  – text analysis
- need whole system to test your contribution

The Festival Speech Synthesis System

Three types of target user

- ☐ Framework for speech synthesis research
- ☐ Component in larger language system
- ☐ Black box text-to-speech system

Festival as software

- ☐ General system for multi-lingual TTS
- ☐ C/C++ code with Scheme scripting language
- ☐ General replaceable modules:
  - lexicons, LTS, duration, intonation, phrasing, POS tagging, tokenizing, diphone/unit selection, signal processing
- ☐ General tools:
  - intonation analysis (F0, Tilt), signal processing, CART building, n-gram, SCFG, WFST, OLS
- ☐ No fixed theories
- ☐ New languages without new C++ code
- ☐ Multiplatform (Unix/NT)
- ☐ Full sources in distribution
- ☐ Free software

Festival as methodology

- individual research can become part of system
- whole system keeps research focused
- whole system keeps commercial interest
- other projects may use results easily
- users' needs help direct the work
- lots of users give lots of feedback

Uses of Festival

- TTS in various domains
- Rapid building of new voices/languages
- Concept to speech systems
- dialog systems
- singing synthesis
- talking heads

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.*

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

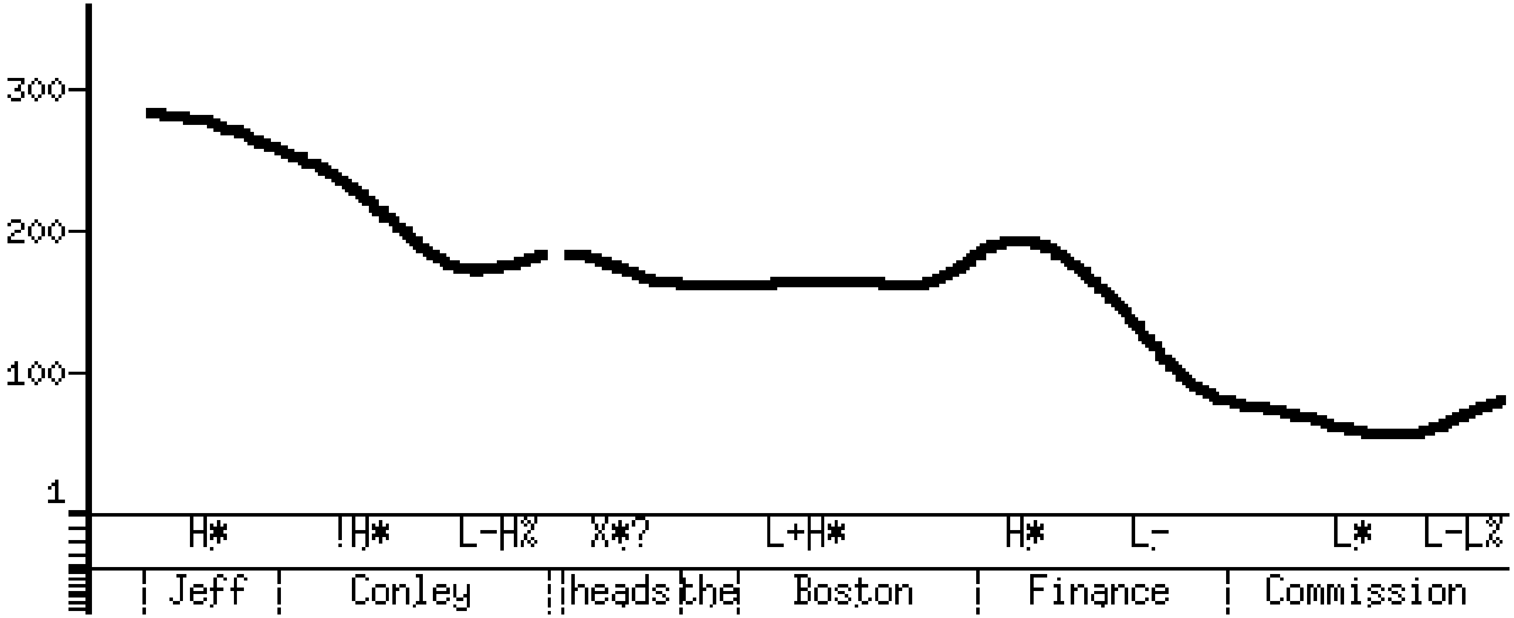
- SABLE is quite limited
- Concept to speech is richer:
  - translation and generation systems
  - Syntactic, Semantic
  - Anaphoric, Rhetorical, Speech act etc.
- Mark up should be:
  - abstract not low-level
  - e.g *type=question* not
  - *pitch rise at end*

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

Prosodic Models

- Style, Speaker characteristics
- Fast/Slow Speech
- Non-natural prosody
- appropriate rendering on semantic markup

Waveform generation

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

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

Acoustic distance

mean weighted Euclidean distance between frames To find most similar units define acoustic distance between two units of the same type U, V

$$Adist(U,V) = \begin{cases} \text{if } |V| > |U| \quad Adist(V,U) \\ \frac{WD*|U|}{|V|} * \sum_{i=1}^{|U|} \sum_{j=1}^{n} \frac{W_j.(abs(F_{ij}(U) - F_{(i*|V|/|U|)j}(V)))}{SD_j * n * |U|} \end{cases}$$

$|U|$ = number of frames in U
$F_{xy}(U)$ = parameter y of frame x of unit U
SD_j = standard deviation of parameter j
W_j = weight for parameter j
WD = duration penalty

Frames include: F_0, 12 MFCC, Energy, delta MFCC

Making clusters

Classification and Regression Trees (Breiman84)

Impurity(Cluster) = mean acoustic distance between members

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

Recursively find best question which splits C such that mean impurity of sub-clusters less than impurity if C.

Questions use:
– phonetic context
– pitch and duration context
– Syllable position, stress, accent
– Position in phrase

i.e. features that exist at synthesis time

```
(w
 ((p.name is #)
  ((duration < 0.0394)
   ((((10 26 31 49 50 55 61 85 89 90 103 233))))
   ((((1 24 86 92 96 124 127 129 131 144 ...)))))
  ((p.name is n)
   ((((2 12 29 59 66 ...))))
   ((n.name is oo)
    ((((5 8 23 30 33 67 ...))))
    ((p.name is @)
     ((n.ph_vheight is 2)
      ((((13 14 106 ...))))
      ...
```

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

Other synthesis uses

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

Directions in speech synthesis

- ☐ Statistical data-driven processes
- ☐ more "ASR" techniques:
  - – language modelling
  - – acoustic modelling
  - – autolabelling
- ☐ Better integration into other systems
  - – not just **text**-to-speech

Synthesis Goal

To make computer synthesized voices as
– natural
– flexible
– efficient
as human voices