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

Target audiences

- Embedded systems: PDA:
  - local rendering of speech on PDA
  - digital radios, books
  - toys and games
  - speech-to-speech translation
- Server systems:
  - multi-channel telephone dialog systems
- But these are "just" commercial uses:
  - No, not exclusively
  - these are also for the speech/dialog researcher
  - when a component can be easily added it will be
  - No excuses for avoiding speech technology

Other uses of speech synthesis

- Speech-to-speech translation
- Talking heads
- Singing

Speech-to-speech translation

- ☐ Multilingual
- ☐ Voice conversion:
  – should speak as source speaker in target language
- ☐ Style/prosody conversion:
  – commands should be commands
  – questions should be questions

Talking heads

- ☐ Adds novelty to conversation
- ☐ Experiments show better understanding:
  - – lip synching
  - – facial movement
- ☐ Listeners swear its better synthesis

Some examples

- ☐ Baldi:
  - – UCSC, Don Massaro, animation
  - – OGI Fstival voice
- ☐ Baphy (CMU):
  - – Off-line mpeg builder
  - – simple faces plus audio alignment
- ☐ JPL/Savante, John Wright:
  - – photo realistic faces
- ☐ Standford/Interval's Video Rewrite:
  - – unit selection type video

How do they work

- ☐ Synthesize audio:
  - – output phone position in audio stream
- ☐ Map phones to faces
- ☐ Build visual stream:
  - – choose appropriate frames
  - – aligned with audio

How many facial positions?

Visemes

- (Baphy) Three positions:
  - closed, open, rounded
- (Cepstral's Rho):
  - 10 lip positions
  - eyelid 4
  - eyes 2
  - other ...
- When should they align:
  - follow trajectories, not fixed times
  - shape for syllable not phone

Synthesis Analogy

Note the analogy to three basic synthesis types

- ☐ Articulatory:
  - modelling the vocal track
  - Baldi: movement of muscles
- ☐ Formant:
  - modelling of signal synthetically
  - cartoon based faces (Baphy)
- ☐ Concatenative:
  - joining natrual segments
  - JPL example
  - Interval's Video Rewrite
- ☐ Unit size:
  - Baphy == uniphone
  - JPL == diphone
  - Video Rewrite == unit selection

Singing

- ☐ Not just speaking but singing too
- ☐ Simple pitch and duration control
- ☐ Proper singing synthesizers:
  – recording singing database
  – singing voice not spoken

Flinger (Festival Singer) OGI/Macon

- Sinusoidal Modelling:
  - uses the OGI diphone dbs
- MIDI interface:
  - allow mixing with music
  - standard MIDI authoring techniques

Festival XML Singer

- Dominic Mazzoni (11-752, project 2001)
- XML based song description:

```
<DURATION BEATS="1.0">
<PITCH NOTE="C4">Oh</PITCH>
</DURATION>
```

- But not just setting pitch duration:
  – when do you move to new pitch
  – how do you move to new pitch

```xml
<?xml version="1.0"?>
<!DOCTYPE SINGING PUBLIC "-//SINGING//DTD SINGING mark up//EN"
      "Singing.v0_1.dtd"
[]>
<SINGING BPM="30">
<PITCH NOTE="G3"><DURATION BEATS="0.3">doe</DURATION></PITCH>
<PITCH NOTE="A3"><DURATION BEATS="0.3">ray</DURATION></PITCH>
<PITCH NOTE="B3"><DURATION BEATS="0.3">me</DURATION></PITCH>
<PITCH NOTE="C4"><DURATION BEATS="0.3">fah</DURATION></PITCH>
<PITCH NOTE="D4"><DURATION BEATS="0.3">sew</DURATION></PITCH>
<PITCH NOTE="E4"><DURATION BEATS="0.3">lah</DURATION></PITCH>
<PITCH NOTE="F#4"><DURATION BEATS="0.3">tee</DURATION></PITCH>
<PITCH NOTE="G4"><DURATION BEATS="0.3">doe</DURATION></PITCH>
</SINGING>
```

Future directions in speech synthesis

- Naturalness:
  – sounding human
- Flexibility:
  – sound the way you want
- Efficient:
  – number of parameters

Some PhD topics ...

Decomposition of Speech

- You need examples to build models:
  - phone, stress, syl position
  - duration, F0, power
  - style etc
- or decompose, build independent models
  - prosodic
  - spectral
  - residual
- Ideally decompose into parameters that are:
  - independent
  - separately trainable
  - (probably) articulatory grounded

Unit selection synthesis

Selecting appropriate units from speech databases

- ☐ What acoustic measures best capture selection
- ☐ What acoustic measures best capture joining
- ☐ How does database size affect quality
- ☐ How can large databases be indexed efficiently
- ☐ Can building databases be more efficient
- ☐ Can databases be interestingly compressed
- ☐ What should the unit size and type be:
  – sub-phonetic, diphone etc
  – acoustically derived units

Prosodic Modelling

- ☐ Better, trainable, stylistic
- ☐ What are the degrees of perception:
  – how can be measure prosody effectively
- ☐ How to model emphasis, contrast
- ☐ How to model sarcasm, politeness
- ☐ Should duration, F0 and power be modelled together

Voice conversion

- ☐ Record everything:
  – or record a little and interpolate
- ☐ Capture speaker characteristics from short examples:
  – spectral, residual and prosody
- ☐ Carefully record core speakers:
  – modify core speaker closest to target

Speech Generation

Concept-to-Speech

- Current language/speech geenration:
  - generate text, plus text to speech
  - generate marked up text, plus text to speech
- Lexical, prosodic choice in generation
- What discourse acts influence speech
- How do we render discourse act information as speech

Information through audio channel

- ☐ Audio channel is limited:
  - – you can't read list of 20 flights
- ☐ How do you best present information:
  - – through dialog
  - – through lexical selection
  - – through prosodic/spectral marking
- ☐ How do you focus to important information:
  - – which part should be focus
- ☐ How do you combine with other modalities:
  - – graphics
  - – tables
  - – talking heads

Information through noisey audio channels

- How is understandability affected when:
  - listener is busy
  - listener is in noisey environment
  - listener is elderly or non-native speaker
- How can you make speech more understandable:
  - in adverse conditions
  - lexicon, prosodic, spectral modification

Fast Speech

- How can you get audio information faster:
  - – audio isn't very fast
  - – blid people play synthesis at 3-5 times real time
- Making it faster:
  - – speeding it up (non-linear)
  - – focussing words
  - – be able to replay and vary speed
  - – skipping and summaries
  - – good dialog/UI models

Spectral modification of units

- ☐ Given limited speech examples (e.g. diphones):
  – modify the spectral properties
  – interpolate from limited examples
- ☐ Do conversion for style:
  – articulation (spectral tilt)
  – shouting, whispering (spectral tilt)
  – joining units by spectral smoothing

Consider similar for prosody

Talking heads and personality

- What makes talking heads more engaging:
  - How can this be used to improve communication
  - How to control facial movements (cf. prosody)
- How to produce personality:
  - SpeechWorks "voice" is engaging
  - can this be synthesized with face/voice efficiently
- Why does it need to be a human face?
  - animal characters,
  - objects (airplanes, ...)
  - gaming industry needs this

Evaluation of synthetic speech

- What is good and bad:
  - how to measure human perception of speech
  - quality, style etc
- Find corelates of human perception:
  - acoustic autmatic measures that model human perception
- Diagnostic evaluation:
  - Which part is good or bad (and their relationship)
  - Why?