



Speech Processing 11-492/18-492

Speech Synthesis

Talking heads

Singing Synthesis

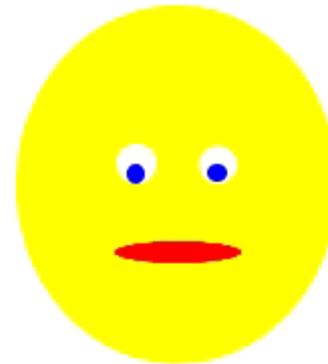
More Information is Better

- ◆ *Voice + text is easier to understand*
- ◆ *Voice + face is easier too*

Talking Heads

- ◆ *Adds novelty/character/personification*
- ◆ *Experiments show better understanding*
 - *Lip synching*
 - *Facial movements*
- ◆ *Listeners swear its better synthesis*

Talking heads



Talking Heads

- ◆ *Synthesize text*
 - *Output phone position in audio stream*
- ◆ *Map phones to lip/tongue positions*
- ◆ *Build visual stream*
 - *Choose appropriate frames*
 - *Aligned with audio*
- ◆ *How many facial positions*

Visemes

- ◆ *Baphy Three positions*
 - *Closed, open and rounded*
- ◆ *Rho*
 - *10 lip positions*
 - *Eyelid 4*
 - *Eyes 2*
- ◆ *When should the align*
 - *Follow trajectories, not just at time instant*
 - *Shape for syllables not just phones*

Synthesis Analogies

- ◆ *Articulatory Synthesis*
 - *Modeling the vocal tract*
 - *Baldi: movement of muscles*
- ◆ *Format:*
 - *Modeling of signal synthetically*
 - *Carton based faces (Baphy)*
- ◆ *Concatenative*
 - *Joining natural segments*
 - *JPL example*
 - *Interval's Video Rewrite*
- ◆ *Unit size*
 - *Baphy == uniphone*
 - *JPL == diphone*
 - *Video Rewrite == unit selection*

Talking Heads

- ◆ *Personalization:*
 - *Can look like a mask put on a dummy*
- ◆ *Uncanny valley*
 - *The more human like, the more critical we are*
- ◆ *3-D movement (in real time)*
 - *Second-life type characters*
 - *Gesture generation too*
- ◆ *Off-line*
 - *(Gollum, Jabba the Hut)*
 - *Usually actors do the voices*

Singing Synthesis

- ◆ *Simple pitch and duration control* 
 - *But singing is more than that*
- ◆ *Proper singing synthesis*
 - *Recording a singing database*
 - ⊗ *Phonetic, prosodic, and singing style coverage*
 - *Sang rather than spoken voice*

Flinger (Festival Singer) (Macon)

- ◆ *Sinusoidal modeling*
 - *More pitch control than just PSOLA*
- ◆ *MIDI interface*
 - *Allow mixing with music*
 - *Standard MIDI authoring techniques*



Festival Singing Mode

- ◆ *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 at duration point*
 - *When do you move it (based on syllable and voicing)*
 - *How quickly do you move pitch*

Singing Example

◆ `<?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>`



Vocaloid

◆ Yamaha: user controlled singing synthesis

The screenshot displays the VOCALOID Editor interface for the file 'korya B.mid'. The main window features a piano roll with a yellow background and a grid. The score is written in a staff with notes and lyrics. The lyrics are: "fun", "its", "so", "sad", "m", "you", "best friends", "the", "are". The notes are color-coded: blue for 'fun', green for 'its', red for 'so', purple for 'sad', yellow for 'm', orange for 'you', pink for 'best friends', light blue for 'the', and dark blue for 'are'. The interface includes a menu bar (File, Edit, View, Job, Track, Lyrics, Setting, Help), a toolbar with various editing tools, and a status bar showing grid settings (1/32), quantize (1/32), tempo (137.00), and other parameters. A 'VOCALOID Icon Palette' window is open on the left, showing various attack, vibrato, and dynamic settings. A 'VOCALOID Mixer' window is open at the bottom right, showing volume levels for seven channels and a master channel. The mixer has sliders for MUTE, SOLO, and volume, with values ranging from -7.6 to 0.0. A 'Brightness' control is visible at the bottom left, with a value of 30 and an interval of 30. The interface also shows a piano keyboard on the left and a 'MEASURE' display at the top.

Future in TTS

- ◆ *More natural voices*
 - *Sound human*
 - *Interact in a human way (not just words)*
- ◆ *More personalization*
 - *Sound like a particular person*
 - *Cross lingual synthesis*
- ◆ *More flexible*
 - *Say it with more feeling*
- ◆ *Realtime voice transformation*
 - *Have an American accent while you speak*

Text to speech process

- ◆ *Text analysis*
 - *From characters to words*
- ◆ *Linguistic analysis*
 - *From words to pronunciations*
- ◆ *Waveform analysis*
 - *From pronunciations to noises*

