A Universal Approach to Synthesizing High Quality Speech and Photo-Real Talking Head

Frank K. Soong

Speech Group
Microsoft Research Asia (MSRA)

2010 Faculty Summit, Shanghai
2010-Oct-19
What is TTS

• Text-to-Speech (TTS)
  – an important part of a voice user interface (VUI) for converting input text into speech

• TTS quality
  – naturalness: sounds like human
  – speaker similarity: sounds like the person to be mimicked
  – intelligible: clear and robust
An HMM-based TTS

Training

Speech Database

Statistical Training

HMMs

Adaptation

Personal Speech

Adaptation

Synthesis

Speech Database

Text

Text Analysis

Trajectory Generation

Speech

Vocoder

Waveform Concatenation
## Two Major Approaches to TTS

<table>
<thead>
<tr>
<th>HMM-based Synthesis</th>
<th>Unit Selection Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Statistically trained</td>
<td>• Waveform segment-based unit selection</td>
</tr>
<tr>
<td></td>
<td>• Natural but with occasional glitches</td>
</tr>
<tr>
<td>• Vocoder speech (smooth, stable, high</td>
<td></td>
</tr>
<tr>
<td>intelligibility)</td>
<td>• Large footprint (whole database)</td>
</tr>
<tr>
<td>• Small footprint (less than 2MB)</td>
<td>• More difficult to modify</td>
</tr>
<tr>
<td>• Easy to modify</td>
<td></td>
</tr>
</tbody>
</table>
TTS Technology Advances

- HMM-based TTS: Statistical and Parametric
  - Refine training process
  - Adjust state boundary
  - Improve V/U decision

- Unit Selection TTS: Rich-context Unit Selection (RUS)
  - KLD candidate selection
  - Maximum NCC criterion
  - Prune-and-search in lattice

KLD: Kullback-Leibler Divergence  V/U: Voiced/Unvoiced  NCC: Normalized Cross-Correlation
New Challenges

• How to render natural speech with high intelligibility
New Challenges

• How to generate natural speech with high intelligibility

No speech is more natural than natural speech.
New Challenges

No speech is more natural than natural speech.

Our solutions

• generating a better trajectory: refining HMM
• rendering natural sounding speech: tiling generated trajectory seamlessly with the best waveform segment samples
HMM-Trajectory Tiling based TTS

-- Synthesis

Label sequence

HMM generated trajectories

“Sausage” of waveform segment candidates
HMM-Trajectory Tiling based TTS -- *Synthesis*

HMM generated trajectories

“Sausage” of waveform segment candidates in speech database

Waveform concatenation
Unit Sausage (Lattice) Construction

• To generate a compact “sausage”
  – Context pruning (same label)
  – Beam pruning with a preset threshold
  – Histogram pruning (# of surviving candidates)
NCC based Search in Sausage and Waveform Concatenation

Maximizing normalized cross-correlation (NCC) to optimize

– spectral similarity
– phase continuity
– concatenation time instants
Demos

• 5 hours British English corpus

• 9 hours Mandarin Chinese corpus

TTS Blizzard Challenge 2010: 1\textsuperscript{st} or 2\textsuperscript{nd} place in naturalness, speaker similarity and intelligibility
What’s next

- To enhance the voice user interface (VUI): add a talking head
Photo-Real Talking Head

- Multi-modal VUI
  - An enhanced, natural user interface from single mode (speech or text) to multi-mode (audio + visual)

- Applications
  - Tele-presence, online chat and gaming
  - Computer Assisted Language Learning (CALL) e.g. Engkoo
Speech Synthesis $\rightarrow$ Visual Synthesis

- Our high quality HTT-based approach to Text-to-Speech (TTS)

![Diagram of Speech Synthesis]

- Speech database
- Clustered HMM
- Parametric speech trajectory
  - HMM-based TTS
  - Guided waveform concatenation
  - HMM Tjaectoty Tiling-based TTS
- Waveform inventory
- Waveform concatenation
  - Concatenative TTS
Speech Synthesis → Visual Synthesis

• Same approach to high quality, **visual speech** synthesis

![Diagram showing the process of speech synthesis.](attachment:image.png)
HMM Trajectory-Guided Photo-Real Talking Head

- Small training set (<30 minutes video recording)
- Fully automatic, data driven, real sample rendering
- Lip-sync with speech
- Natural head motion and facial expressions
HMM-Guided Lips Image Selection

HMM-based Visual Synthesis

Image Candidates
Synthesized Lips Movements

• HMM-based vs. HMM-guided

• Summary
  – Intelligible, lip sync, and photo realistic
  – Stitching lips images back to the full face, seamlessly 😊
Speech-to-Lips Conversion

Speech + Phoneme labels + Timing

Speech Only

No. 1 in A/V Consistency Test, LIPS Challenge 2009
Text Driven Talking Head

TTS Voice: Synthesized by RUS
demo
Hello everyone, I am Matt Scott, a photo-real talking
Tele-presence and multi-party gaming application

- High quality speech-to-lips conversion without knowing the underlying linguistic content
- The most presentable face for tele-presence
- Personal choice of talking head in multi-party gaming

Select your favorite head

- [Image of a head]
- [Image of a head]
- [Image of a head]
Summary

• Applications
  – Typical TTS applications, e.g. reading email, news, car navigation,
  – Computer Assisted Language Learning (e.g. Engkoo)
  – Tele-presence and gaming

• Our solutions
  – Statistical modeling and real sample rendering
  – HMM Trajectory Tiling (HTT)-based TTS
  – Photo-real talking head
  – Text or speech driven