Microsoft Research
微软亚洲研究院

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

Frank K. Soong
宋謌平

Speech Group
Microsoft Research Asia (MSRA)

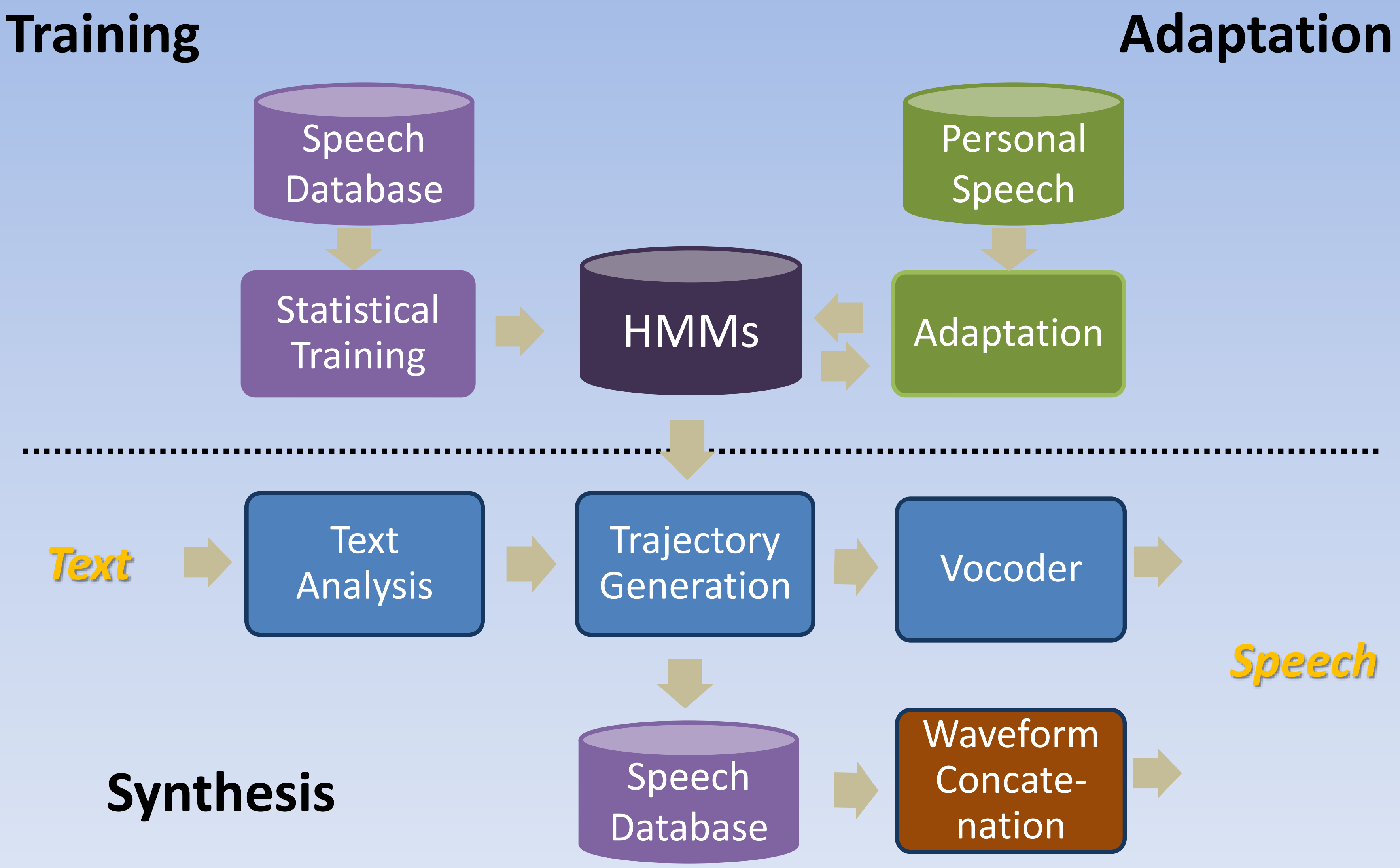
2010 Faculty Summit, Shanghai

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

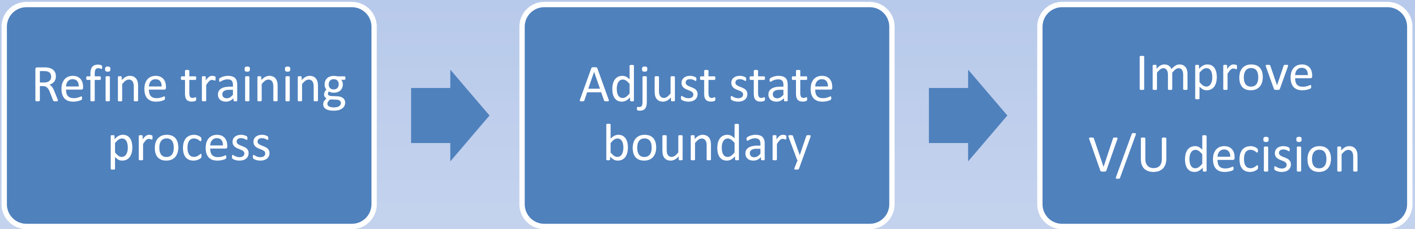


Two Major Approaches to TTS

HMM-based Synthesis	Unit Selection Synthesis
• Statistically trained	• Waveform segment-based unit selection
• Vocoded speech (smooth, stable, high intelligibility)	• Natural but with occasional glitches
• Small footprint (less than 2MB)	• Large footprint (whole database)
• Easy to modify	• More difficult to modify

TTS Technology Advances

- HMM-based TTS : Statistical and Parametric

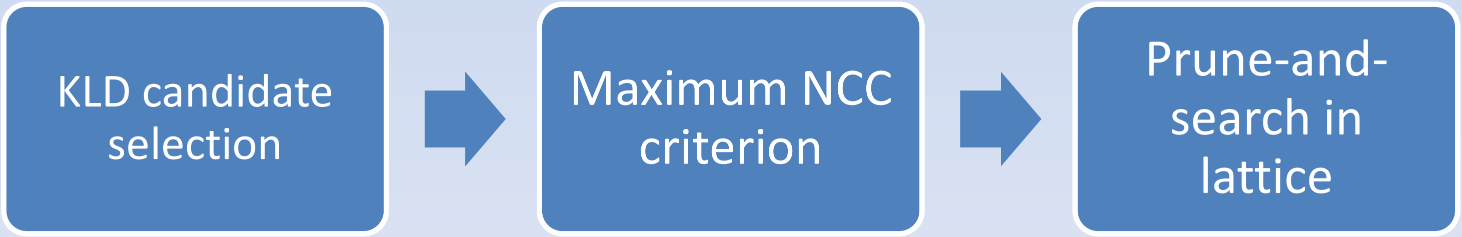


Refine training process

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
  int

No speech is more natural than natural speech.

New Challenges

No speech is more natural than natural speech.

Our solutions

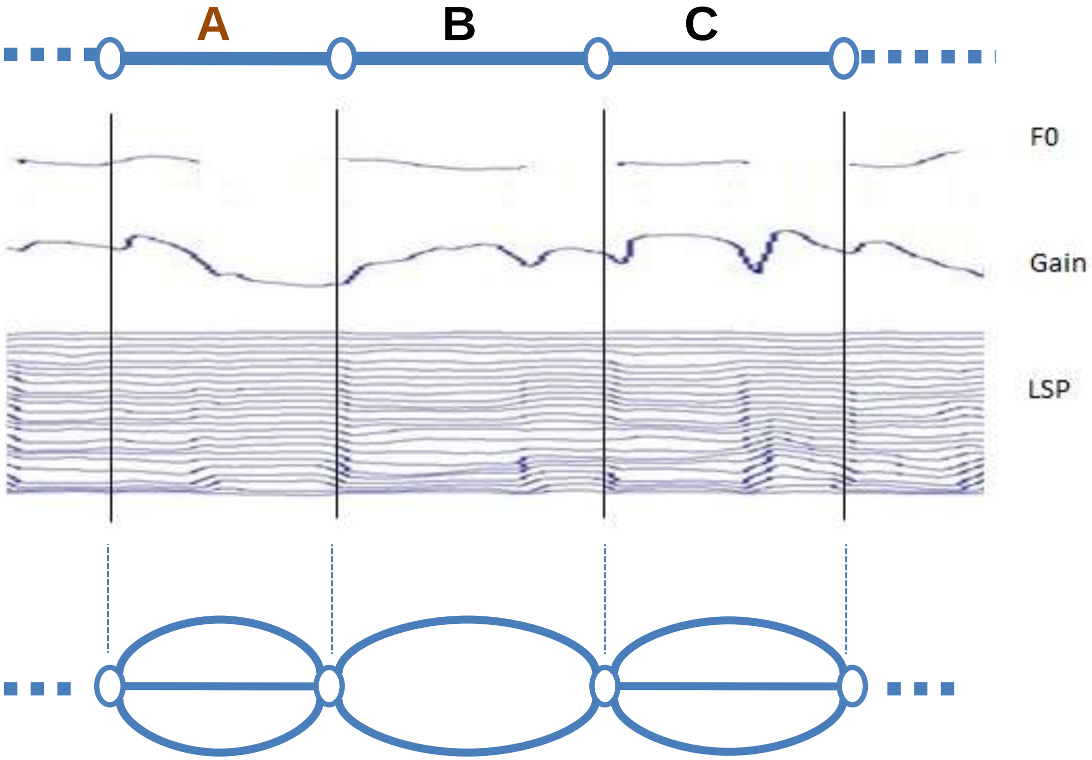
- generating a better trajectory: refining HMM
- rendering natural sounding speech: tiling generated trajectory seamlessly with the best waveform segment samples

HMM-Trajectory Tiling based TTS -- *Synthesis*

Label sequence

HMM generated trajectories

“Sausage” of waveform segment candidates

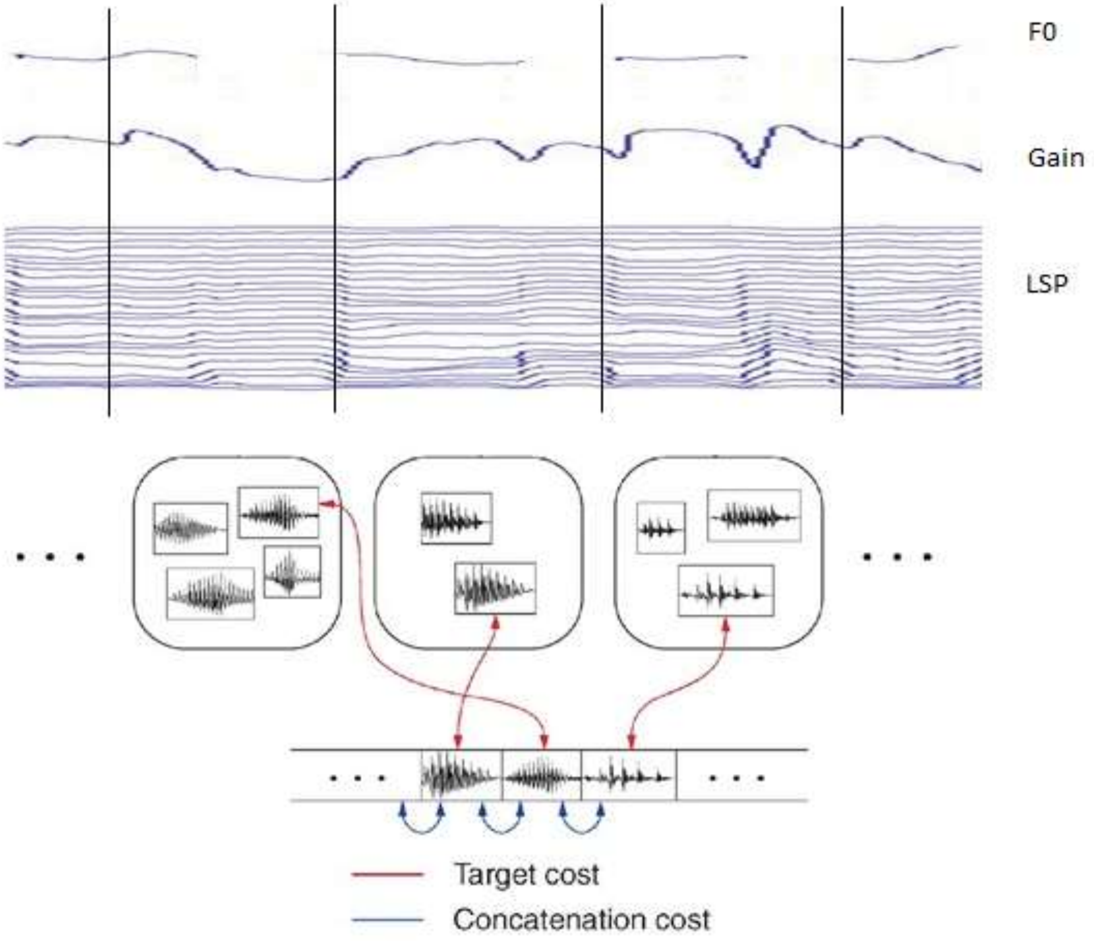


HMM-Trajectory Tiling based TTS -- *Synthesis*

HMM generated trajectories

“Sausage” of waveform segment candidates in speech database

Waveform concatenation



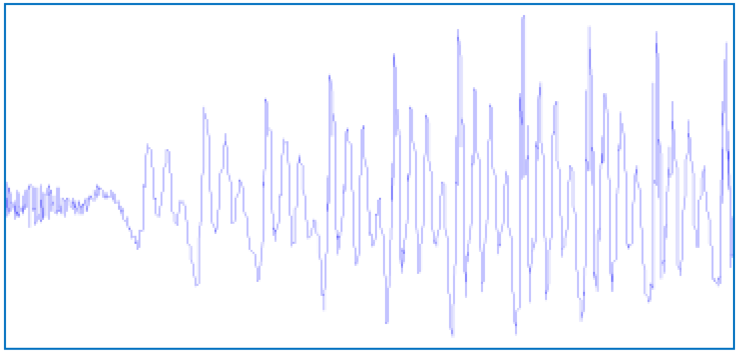
Unit Sausage (Lattice) Construction

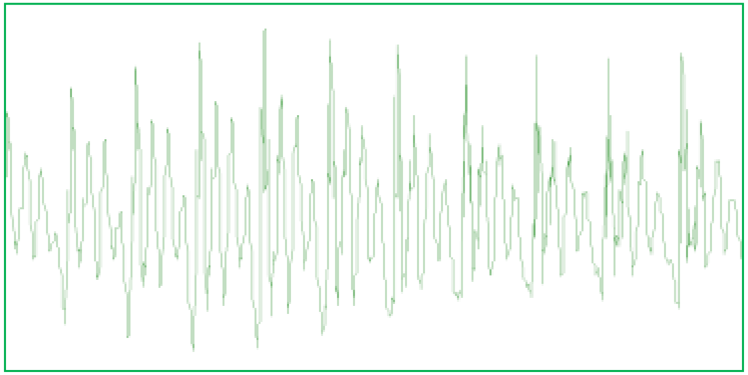
- To generate a compact “sausage”
  - Context pruning (same label)
  - Beam pruning with a preset threshold
  - Histogram pruning (# of surviving candidates)

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

🔈 🔈 🔈 🔈 🔈

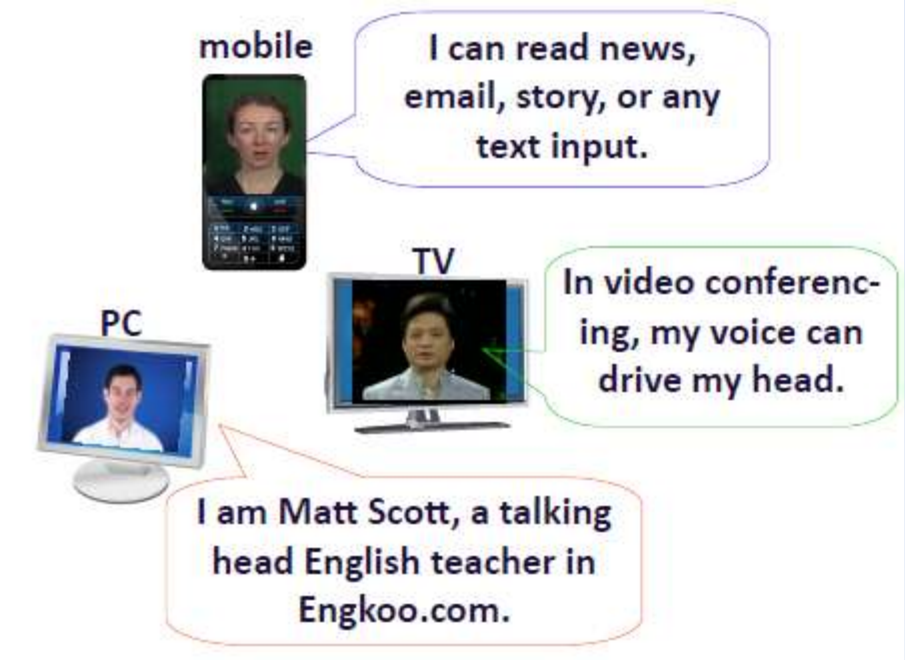
TTS Blizzard Challenge 2010: 1st or 2nd place in naturalness, speaker similarity and intelligibility

What’s next

- To enhance the voice user interface (VUI): add a talking head

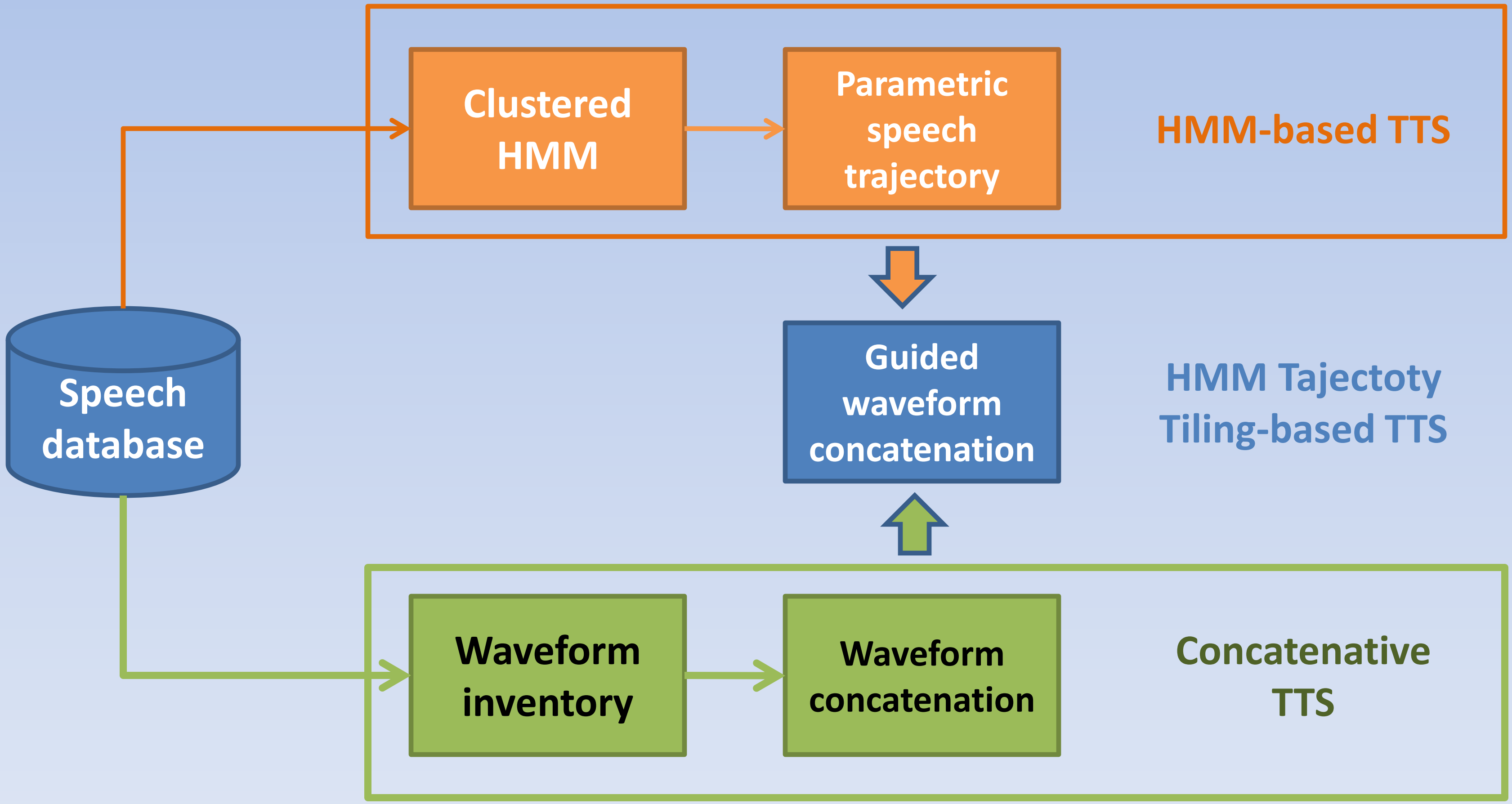
Photo-Real Talking Head

- 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


mobile
I can read news, email, story, or any text input.
TV
In video conferencing, my voice can drive my head.
PC
I am Matt Scott, a talking head English teacher in Engkoo.com.

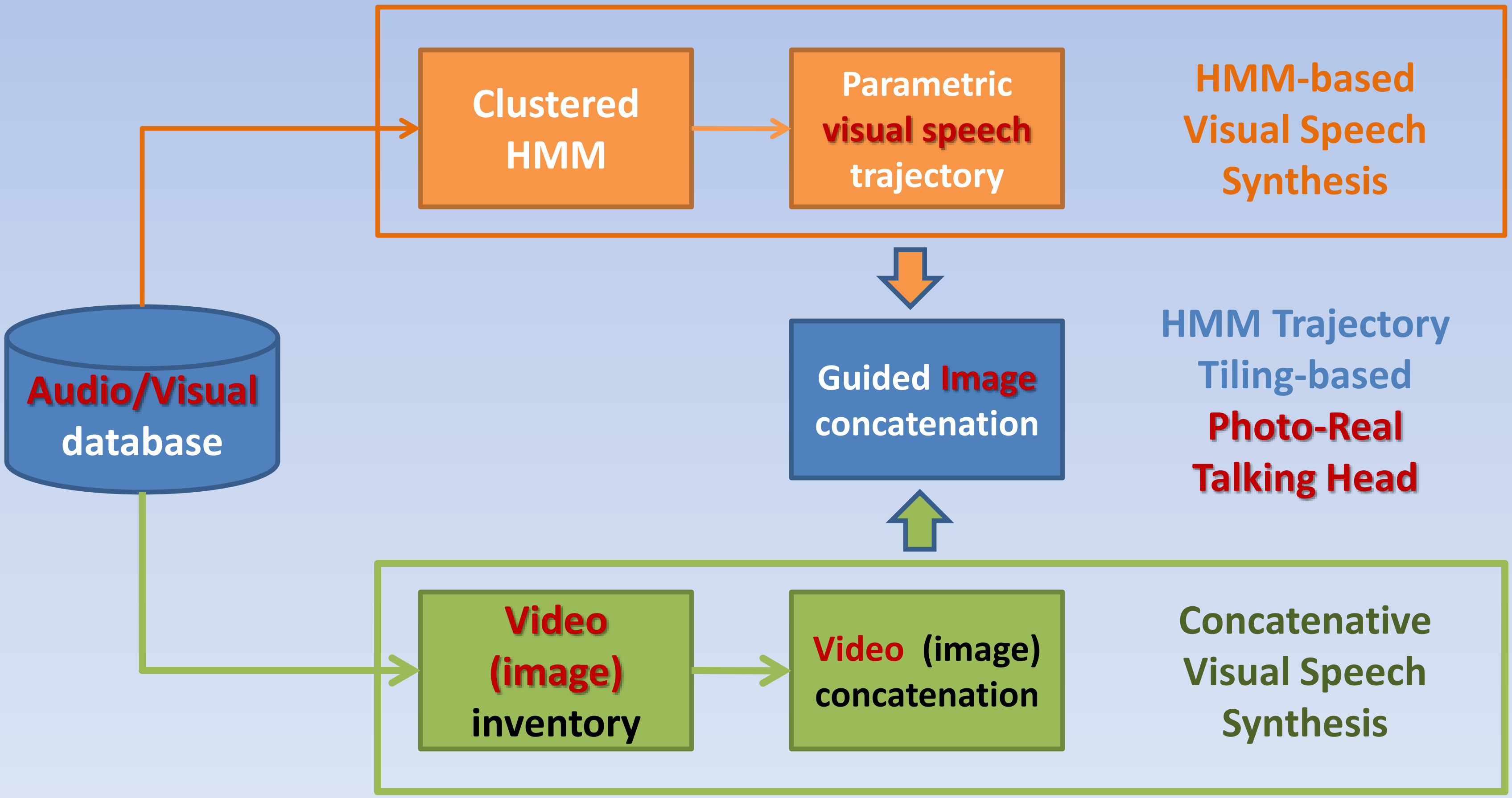
Speech Synthesis ⇨ Visual Synthesis

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



Speech Synthesis ⇨ Visual Synthesis

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



Clustered HMM

Parametric visual speech trajectory

HMM-based Visual Speech Synthesis

Audio/Visual database

Guided Image concatenation

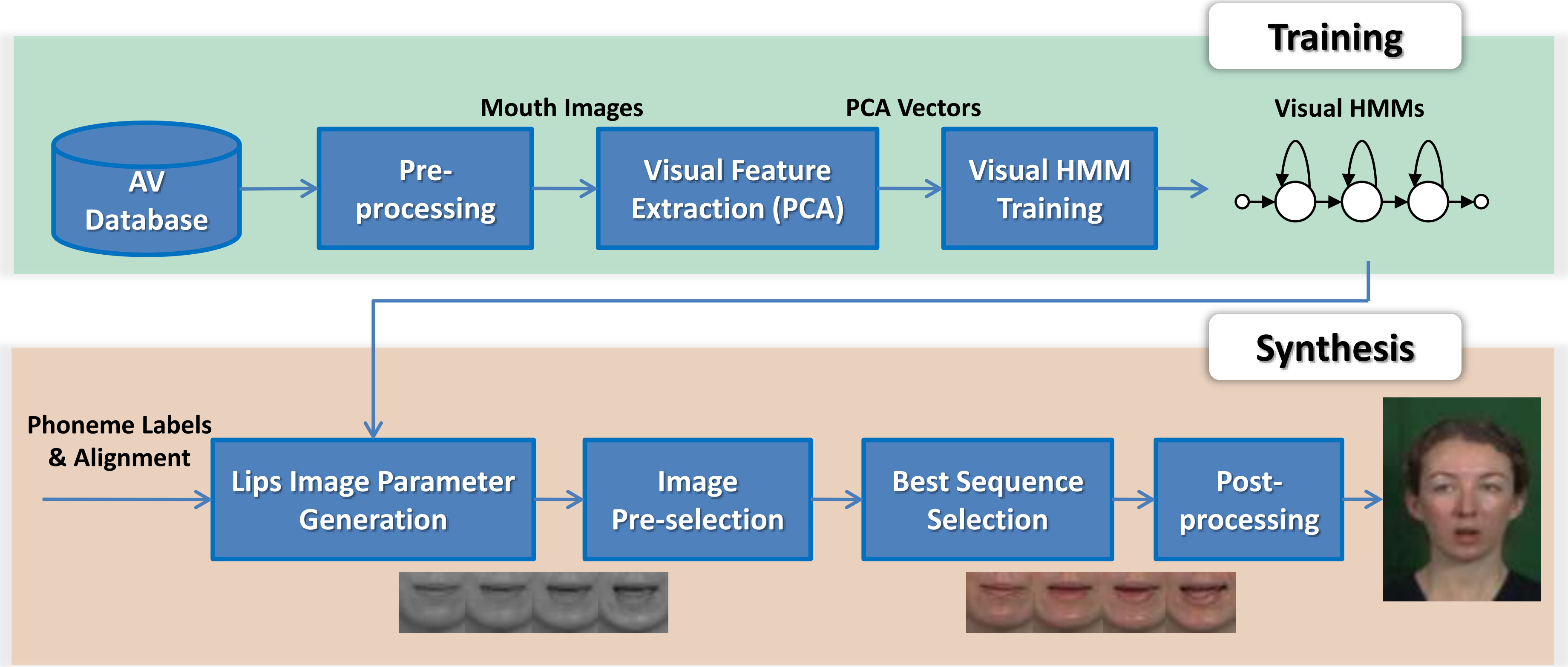
HMM Trajectory Tiling-based **Photo-Real Talking Head**

Video (image) inventory

Video (image) concatenation

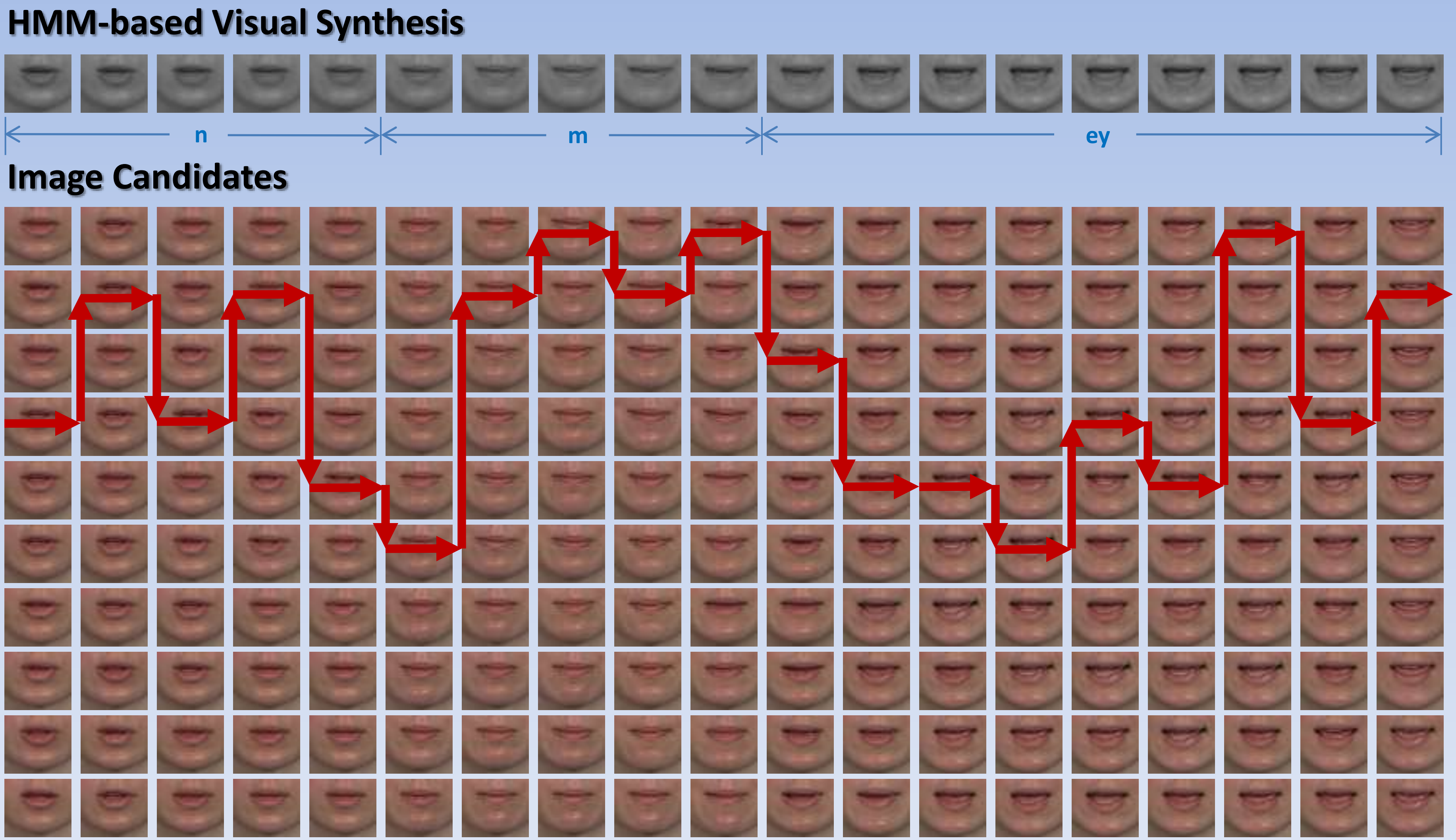
Concatenative Visual Speech Synthesis

HMM Trajectory-Guided Photo-Real Talking Head



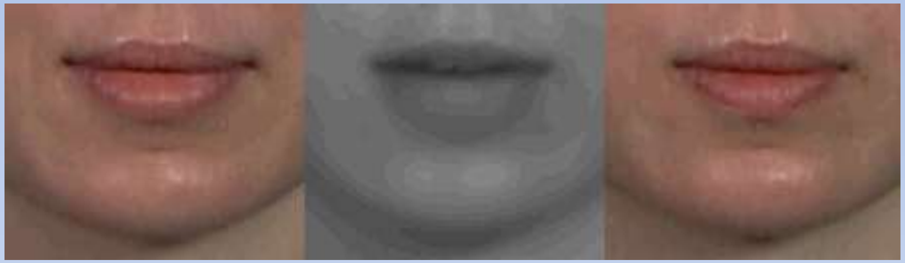
- ✓ Small training set (<30 minutes video recording)
- ✓ Fully automatic, data driven, real sample rendering
- ✓ Lip-sync with speech
- ✓ Natural head motion and facial expressions

HMM-Guided Lips Image Selection

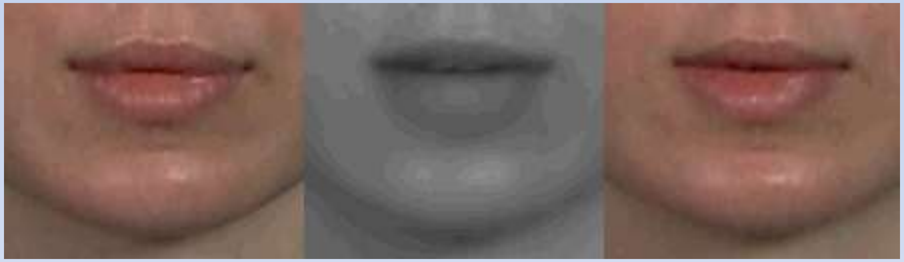


Synthesized Lips Movements

- **HMM-based vs. HMM-guided**



original HMM-based HMM-guided



original HMM-based 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

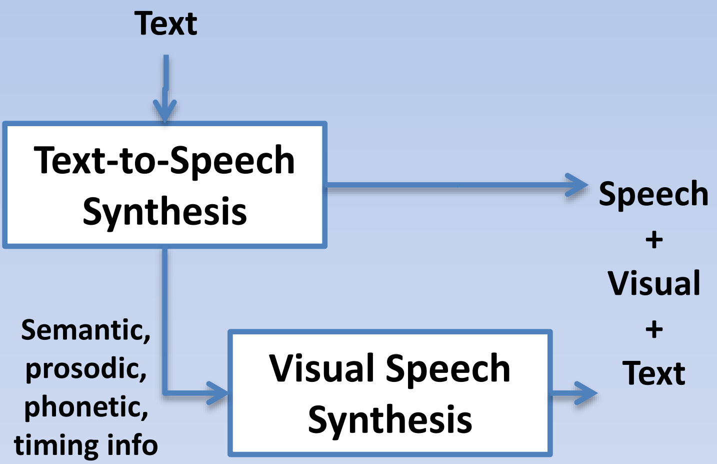
Text Driven Talking Head

TTS Voice: Synthesized by RUS

demo

English Teacher on Microsoft Engkoo 英库

Hello everyone, I am Matt Scott, a photo-real talking



Text

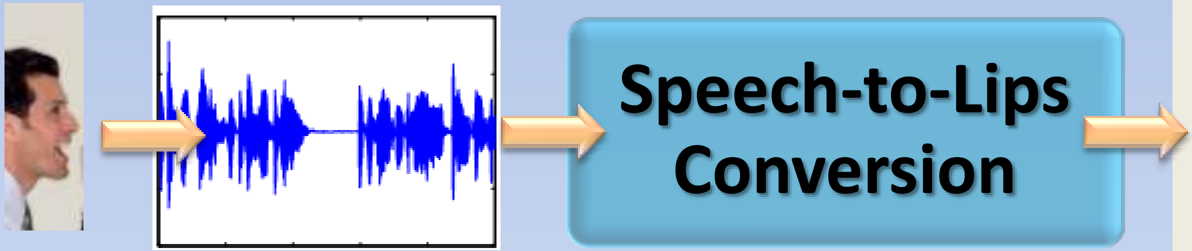
Text-to-Speech Synthesis

Semantic, prosodic, phonetic, timing info

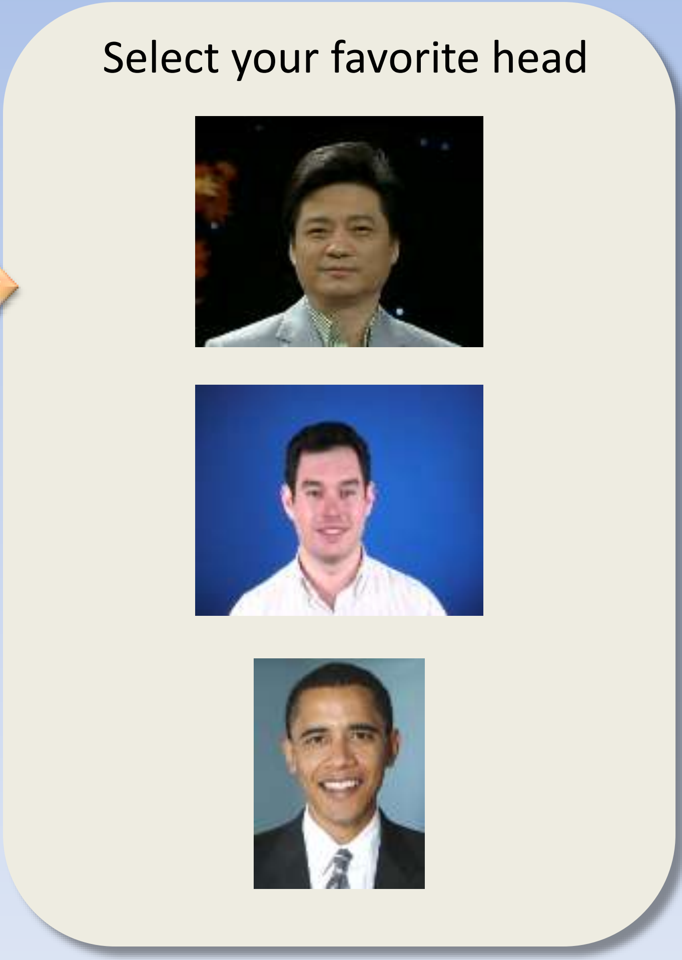
Visual Speech Synthesis

Speech + Visual + Text

Tele-presence and multi-party gaming application



Speech-to-Lips Conversion



Select your favorite head

- 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

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