Special Module on Media Processing and Communication

Dayalbagh Educational Institute (DEI)
Dayalbagh Agra

PHM 961

Indian Institute of Technology Delhi (IITD)
New Delhi

SIV 864

Text-to-Audiovisual Speech Synthesizer

Introduction

- Basic motivation is to add visual channel to text to speech (TTS) interface.
- Various applications

  Face to face communication for Human Computer Interface

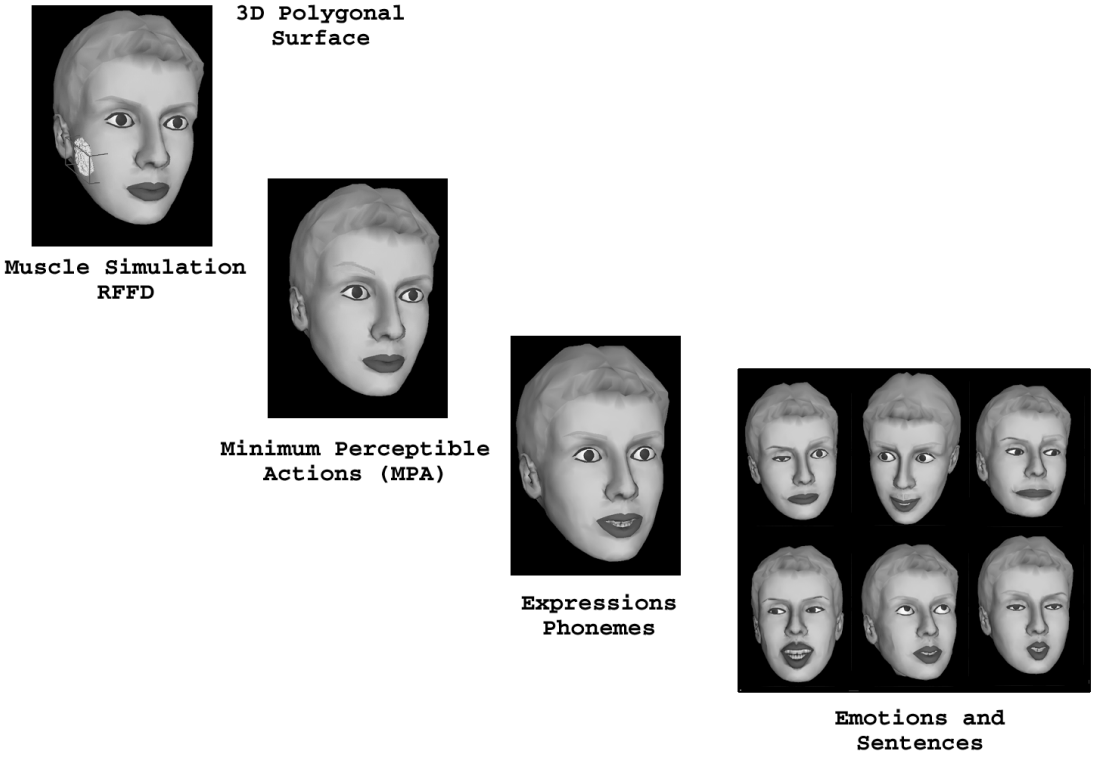
  Video Telephony, video conferencing

Background

- Two approaches
  - Model based
    - Flexible
    - Lacks video realism
  - Image based

Background

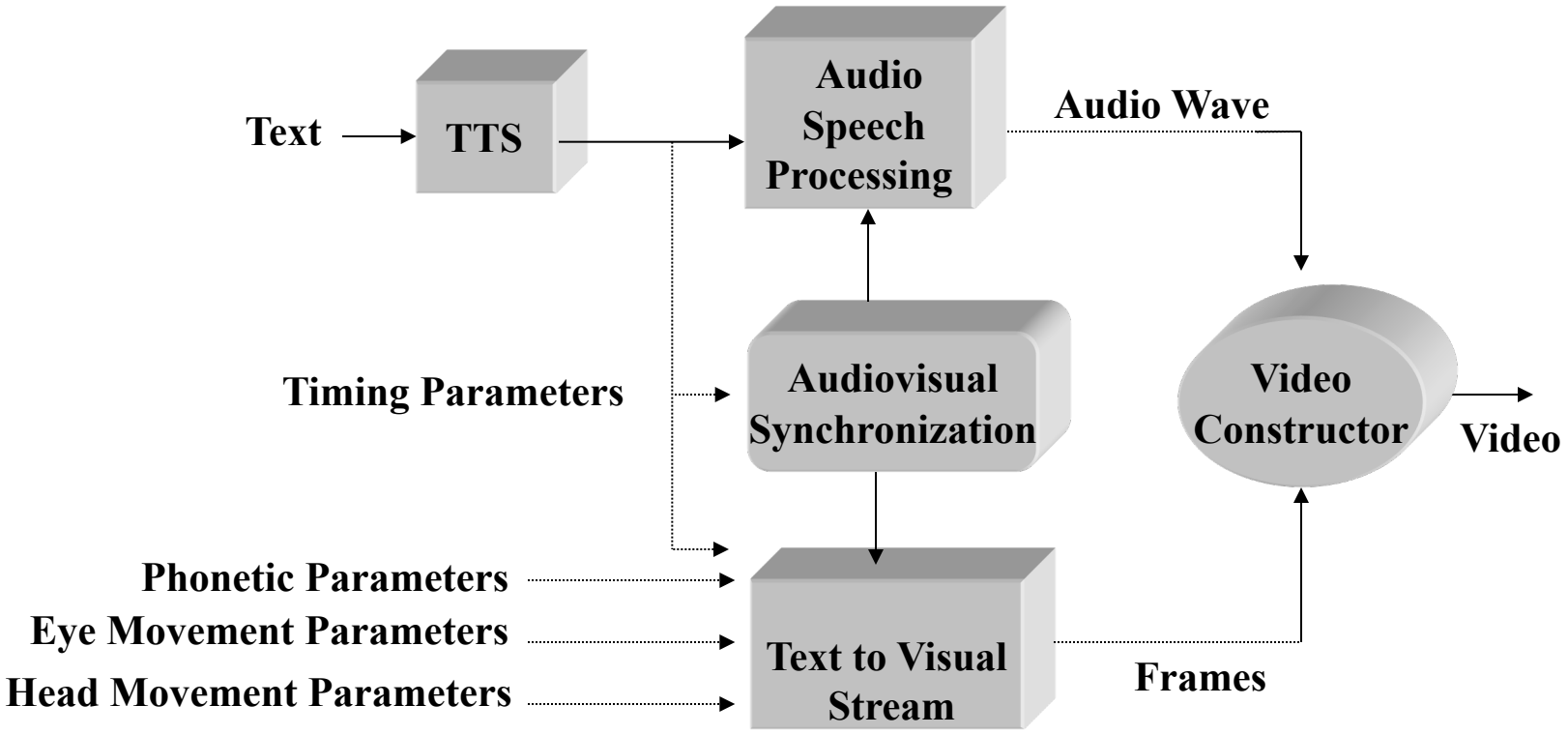
Model based



Approach

- Basic idea
  - Morphing visemes (Ezzat and Poggio)
- Add ons
  - Eye movements
  - Head movements
  - *Co-articulation*

Overview


Text
TTS
Audio Speech Processing
Audio Wave
Timing Parameters
Audiovisual Synchronization
Video Constructor
Video
Phonetic Parameters
Eye Movement Parameters
Head Movement Parameters
Text to Visual Stream
Frames

Text-to-Audiovisual Stream

- Four sub-tasks
  - Viseme Extraction
  - Morphing
  - Morph Concatenation
  - Synchronization

Text-to-Audiovisual Stream

- Viseme Extraction

  Phoneme counterparts

  Many to one mapping

  - 16 snapshots

  Set of keywords covering all the phonemes

/p, b, m/

/oo/

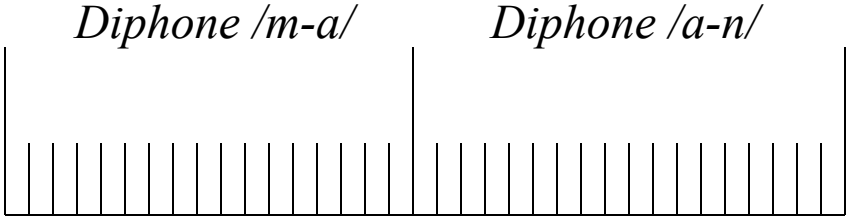
Text-to-Audiovisual Stream

- Morphing
  - Optical flow (automatic correspondence)
  - No other moving part/object (assumption)
- Morph Concatenation
  - Simple concatenation of viseme morphs

Text-to-Audiovisual Stream

- Audiovisual Synchronization

Audio



Viseme Transition

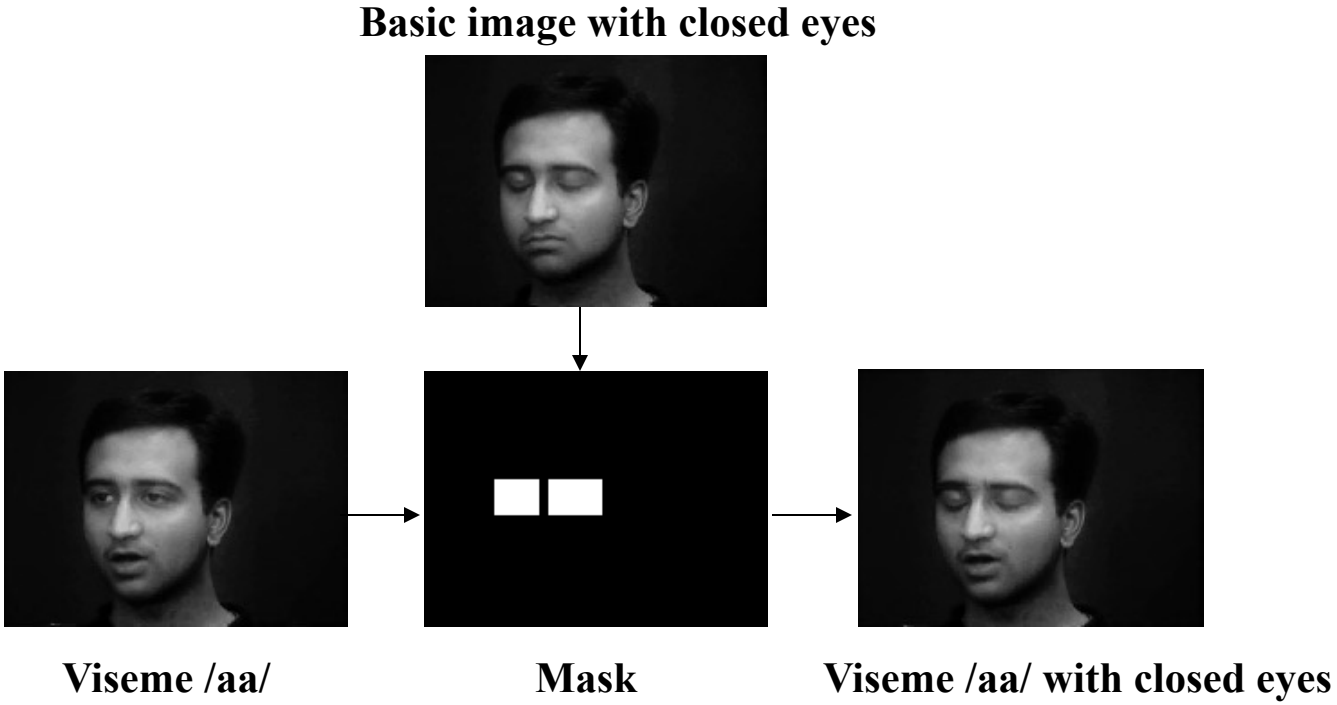
$$\text{Morph parameter} = \frac{s(F_k) - s(V_i)}{I(V_i)}$$

Text-to-Audiovisual Stream

- Examples

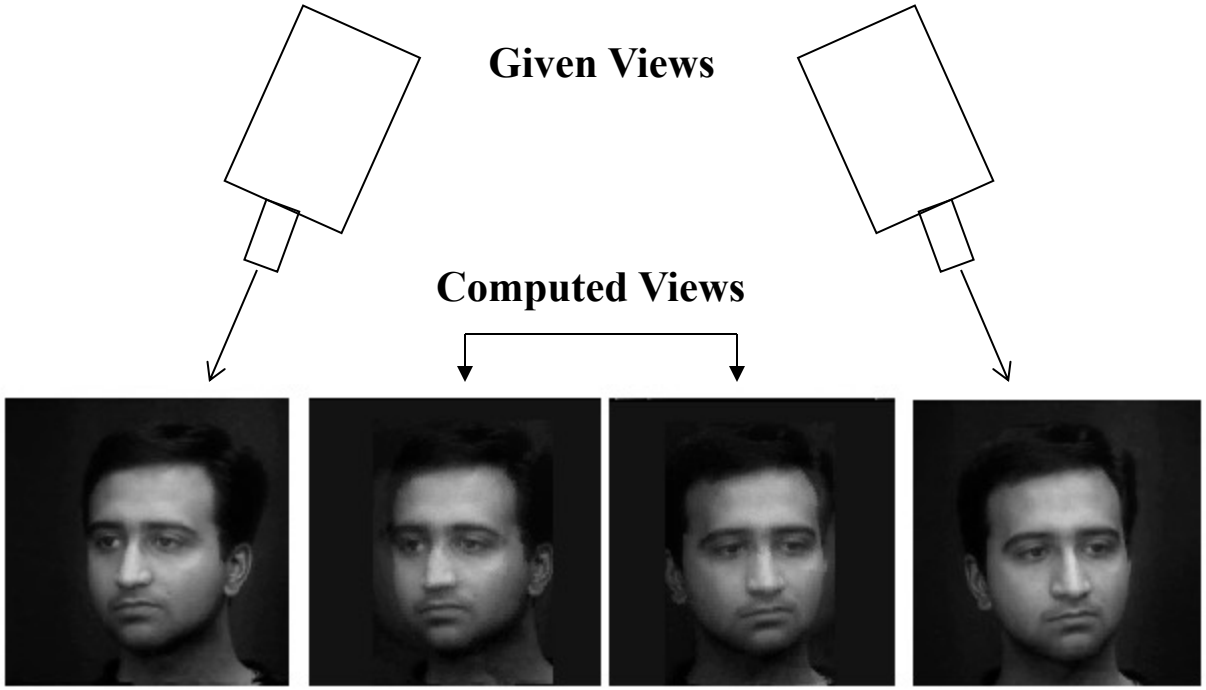
Eye Movement

- Mask based approach

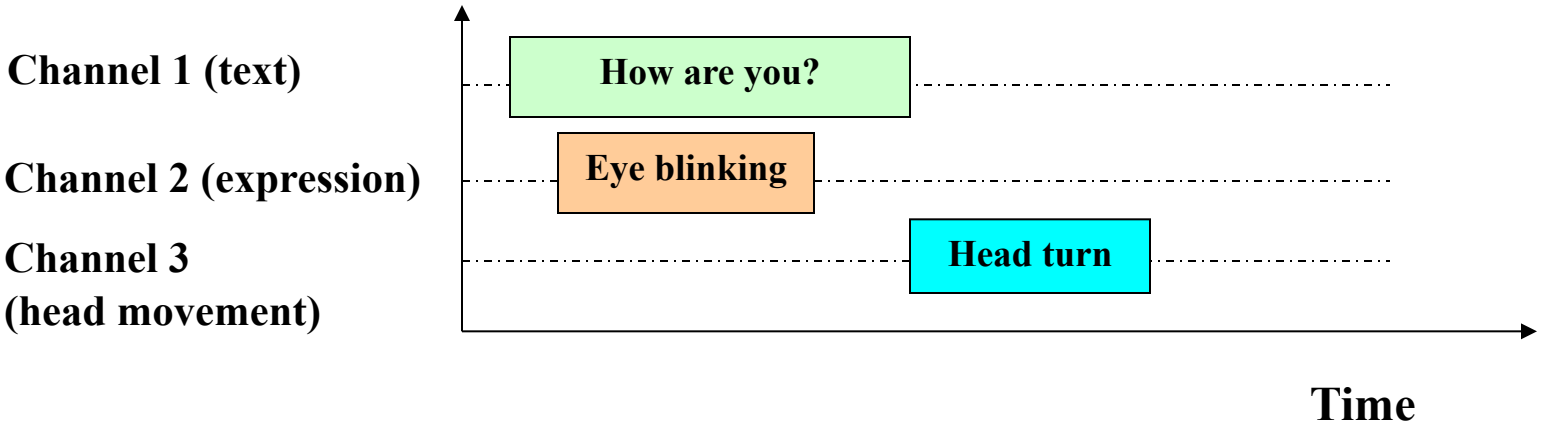


Head Movement

- View Morphing (Seitz and Dyer 1996)



Integration



Results

Demo 1

“I miss you”

Demo 2

“I am fine, thank you”

Co-articulation

- Problem

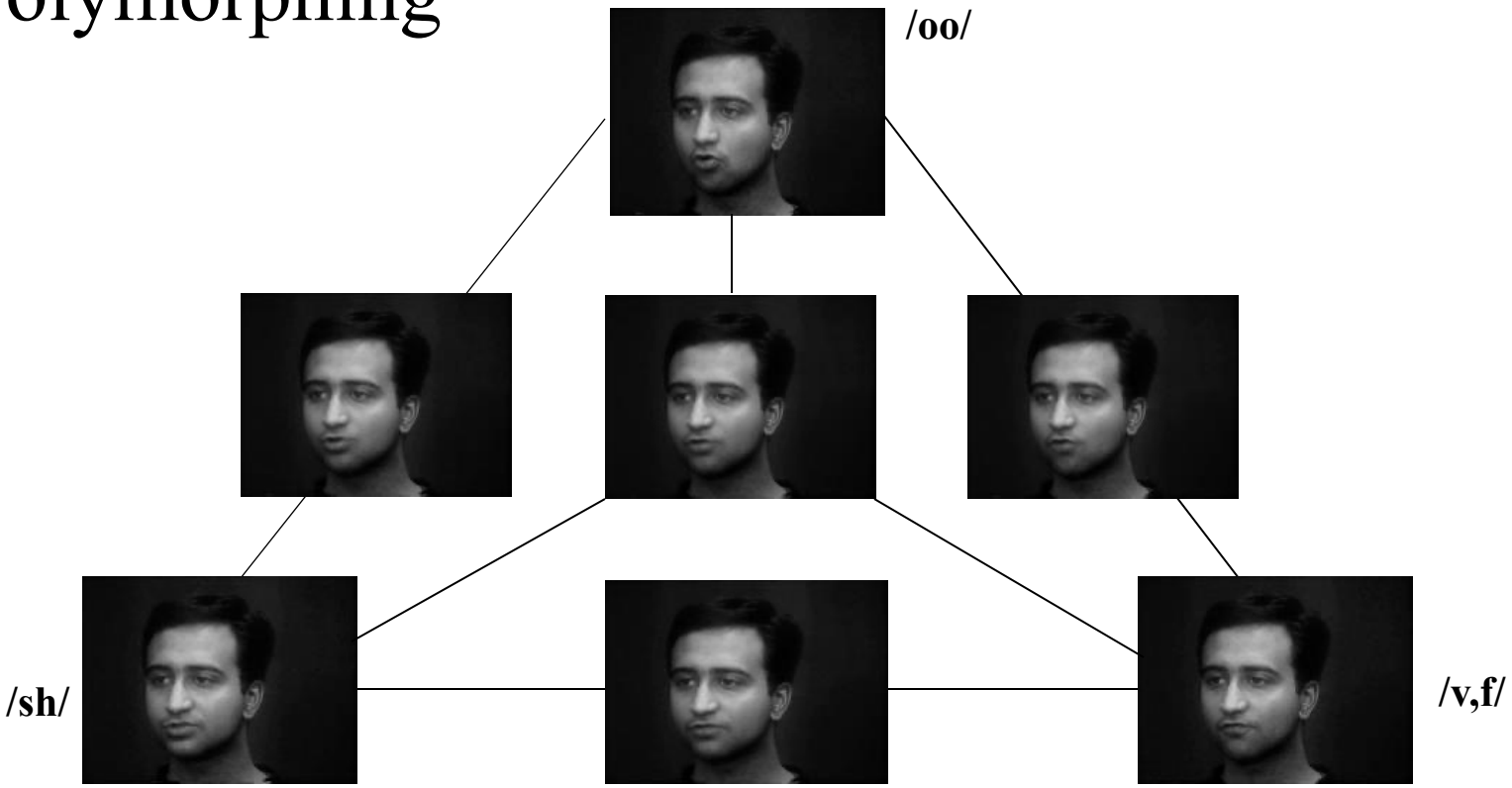
  There is an overlap in the production of syllables and phonemes

- Approach

  Polymorphing

Co-articulation

Polymorphing



Co-articulation

- Some Results

Without co-articulation

With co-articulation

“Tea twenty two temporary food stew”

Conclusion

- A text-to-audiovisual speech synthesizer
- Non verbal communication (expressions and head movement)
- Co-articulation

- Learning based methods