Concatenative articulatory video synthesis of real-time MRI data for spoken language training

Urvish Desai$^1$, Chiranjeevi Yarra$^2$, Prasanta Kumar Ghosh$^2$

$^1$Indian Institute of Technology (ISM), Dhanbad, India
$^2$SPIRE LAB, Electrical Engineering, Indian Institute of Science (IISc), Bengaluru, India.

20th April 2018
Outline

1. Introduction and Motivation
   Discuss the benefit of an articulatory video in the training and need for the proposed work.

2. Proposed approach
   To develop a method for synthesizing the articulatory videos.

3. Evaluation
   Experimental evaluation of the proposed method
Nativity

- Articulatory movements during speaking English are dominated by the articulatory constraints from the speaker’s native language.
- An incorrect phoneme articulation would result in miscommunication.
Importance of Visual training:

- L2 learners would benefit from a video that shows the correct movements of the articulators [1,2].

**Typical approach:**

- Expert’s movements are captured using real time motion capture techniques simultaneously with their audio


Problem Statement

Synthesize an articulatory video corresponding to an expert’s audio for which direct articulatory measurements are not available.
Data acquisition:

- **Electromagnetic Articulography (EMA)**: Lacks complete view, Disrupts speech
- **Ultrasound Imaging**: All the articulators are not visible in a single modality
- **Real-time Magnetic Resonance Imaging (rt-MRI)**: Easy to observe Articulators directly
- **Computed Tomography (CT)**: Causes exposure to radiation


Limitations:

- Data acquisition with these methods is time consuming and expensive.
- Hence, it is challenging to obtain an articulatory video for arbitrary stimuli.
- Typically, stimuli vary across the training methodologies.

Data acquisition:

- Electromagnetic Articulography (EMA)
  - Lacks complete view
  - Disrupts speech

- Ultrasound Imaging
  - All the articulators are not visible in a single modality

- Real-time Magnetic Resonance Imaging (rt-MRI)
  - Easy to observe
  - Articulators directly


**Outline**

- Introduction
- Problem
- Motivation
- **Approach**
- Database
- Evaluation
- Conclusion

**PSIFS repository**

**PSIFS Selection**

**Video Creation**

**Current Workflow**

- Expert Audio
- Forced Alignment
- PSIFS selection using DP
- Selected PSIFS
- Interpolation
- Boundary stitching
- Video creation
- Synthesized video

**Real-time MRI data**

**Video creation**

**Synthesized video**
/ih/ /f/ /d/ /n/ /jh/ /aw/ /th/ 

/w/ /er/ /hh/ /m/ /ao/ /w/ /ae/ 

/ah/ /y/ /ay/ /ah/ /b/ /ah/ /ah/ 

/s/ /z/ /r/ /l/ /eh/ /aa/ /k/ /l/ /v/ /uw/ /sil/ 

/dh/ /iy/ /g/ /ow/ /t/ /ng/ /ey/ /p/ 

PSIFS repository

Phoneme Specific Image Frame Sequence (PSIFS)
• Select such that the Frobenious norm between the last IF of the previous phoneme and the first IF of the next phoneme should be minimum.

<table>
<thead>
<tr>
<th>Mono-phones</th>
<th>/s/</th>
<th>/eh/</th>
<th>/l/</th>
<th>/sil/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrences</td>
<td><img src="image1" alt="Occurrences" /></td>
<td><img src="image2" alt="Occurrences" /></td>
<td><img src="image3" alt="Occurrences" /></td>
<td><img src="image4" alt="Occurrences" /></td>
</tr>
</tbody>
</table>

Based on the phoneme duration $\hat{N}$ is computed.
- Pixel by Pixel linear interpolation of Image frames is performed.
- Two boundary IFs of two consecutive phonemes are merged to obtain one boundary IF by Pixel by Pixel averaging.
MRI-TIMIT database

- 23.18 fps
- 68x68 pixels
- Greyscale
- 20khz sampling frequency
- 2 male and 2 female speakers out of which one is chosen
- 460 TIMIT sentences

Phonetic transcriptions:
Audio is extracted from the videos
Forced alignment using Kaldi SR toolkit (DNN)
Combined lexicon by CMU and TIMIT

40 unique mono-phones

Database

#1: taxi cab broke down

#2: was easy for

#K: with understanding alleviates

#N: worry over silly

#460: the state of

Data for the PSIFS repository

Test Stimuli

taxi cab broke down
with understanding alleviates

the state of
MATLAB GUI

Word string of the current phrase

- **Poor**: There is a great difference between the quality of the synthesized and the original videos. Score is 1.
- **Fair**: There is a moderate difference between the quality of synthesized and the original videos. Score is 2.
- **Good**: There is a slight difference between the quality of synthesized and the original videos. Score is 3.
- **Very good**: There is no significant difference between the quality of synthesized and the original videos. Score is 4.
- **Excellent**: There is no difference between the quality of synthesized and the original videos. Score is 5.
• Averaging the ratings across all the stimuli and all the evaluators, the quality, the quality of the synthesized videos is found to be **3.78 ± 1.07**.
• This indicates the quality of the synthesized video is **not significantly different** from the original video.

### Word: cab-driver Broke down
- Number of Phonemes: 4
- Average rating: 4.36

### Word: XXX Understanding YYY
- Number of Phonemes: 12
- Average rating: 3.09
• An average of **5.7** number of phonemes per word is above 3.78 and **6.8** number of phonemes per word is below 3.78.
• The word **containing more phonemes have more boundaries to smooth** and, hence, could result in **more disruptions** in the synthesized videos.

• However, the ratings do not vary proportionally with number of phonemes in a word.
Conclusion

• We propose a method to synthesize an articulatory video for an audio, for which the articulatory data is not available.
• The proposed method, is based on concatenative synthesis approach, in which, a PSIFS repository is created for every phoneme in the training data.
• Given an audio, we find the best representative PSIFS for each phoneme in a given context to maintain smoothness across the boundaries.
• Following this, we synchronize each selected PSIFS with its respective audio and apply image stitching at the PSIFS boundaries.
• Experiments with MRI-TIMIT containing rt-MRI videos, following subjective evaluation, reveal that the quality of the synthesized video is close to that of the original video.

Future Work

• Further investigations are required to develop better techniques for image stitching as well as for PSIFS selection and interpolation.
• It is also required to propose an objective measure for the evaluation.
Acknowledgement

- We would like to thank all the 12 evaluators who are involved in the subjective evaluation.
- We also thank the Pratiksha Trust for their support
Thank you