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

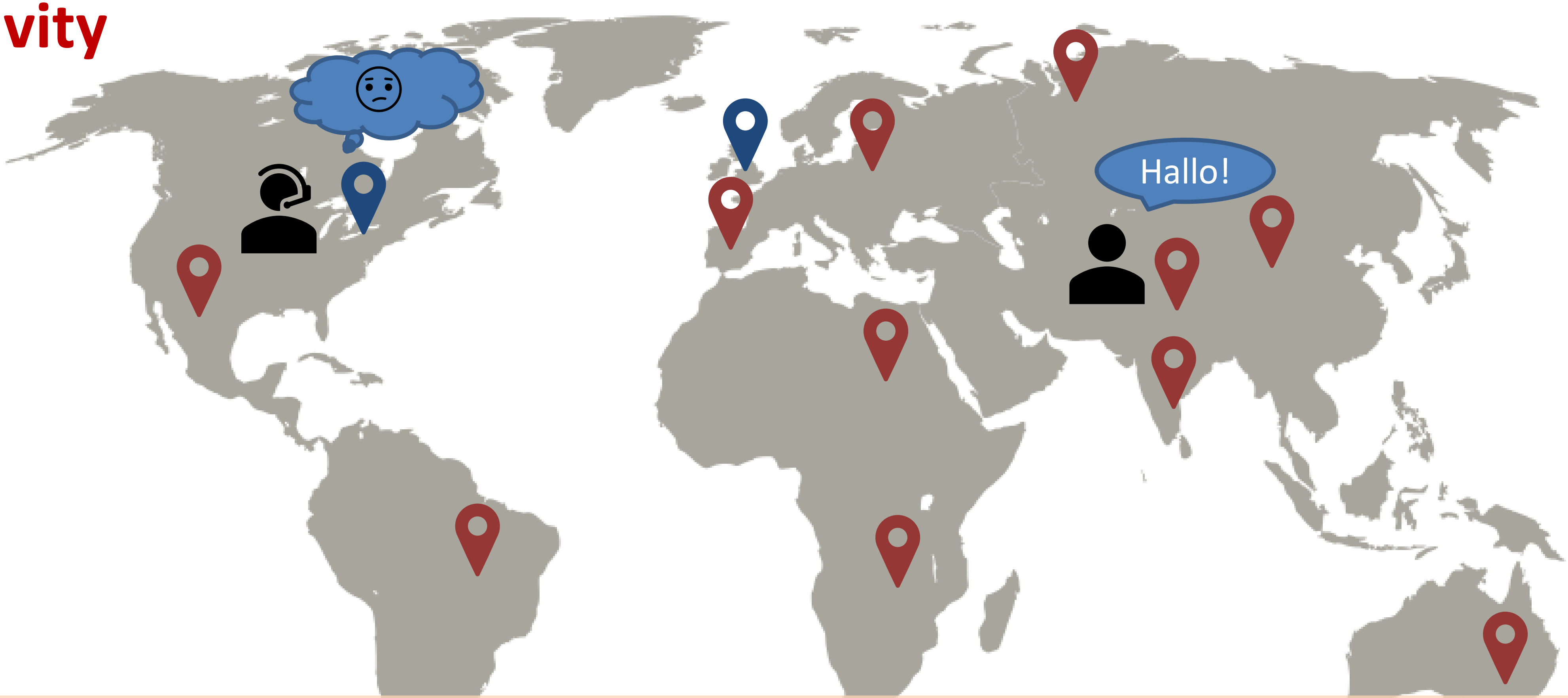
To develop a method for synthesizing the articulatory videos.

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].

Good

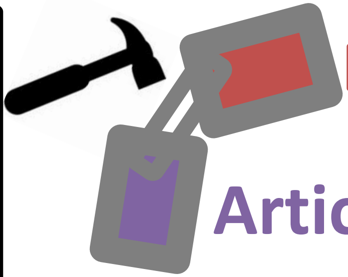
Better

The best

[1] Maxine Eskenazi, "An overview of spoken language technology for education," Speech Communication, vol. 51, no. 10, pp. 832–844, 2009.
[2] Pierre Badin, Yuliya Tarabalka, Fr´ed´eric Elisei, and G´erard Bailly, "Can you 'read' tongue movements? evaluation of the contribution of tongue display to speech understanding," Speech Communication, vol. 52, no. 6, pp. 493–503, 2010.
[3] "How to use Saundz English Pronunciation Software" available at https://www.youtube.com/watch?v=9rUw3wNPJxs

Typical approach:

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



Expert’s Audio

Articulatory movements

[1] Maxine Eskenazi, “An overview of spoken language technology for education,” Speech Communication, vol. 51, no. 10, pp. 832–844, 2009.
[4] Pierre Badin, Atef Ben Youssef, G´erard Bailly, Fr´ed´eric Elisei, and Thomas Hueber, “Visual articulatory feedback for phonetic correction in second language learning,” Workshop on Second Language Studies: Acquisition, Learning, Education and Technology (L2SW), pp. 1–10, 2010.

Problem Statement

Synthesize an articulatory video corresponding to an expert's audio for which direct articulatory measurements are not available.

Expert's Audio → Concatenation based synthesis → Articulatory video

Articulatory database → Concatenation based synthesis

Articulatory database ➔ rt-MRI data

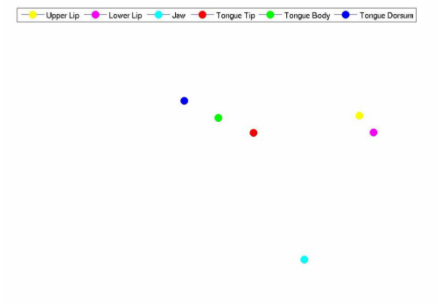
Data acquisition:

Electromagnetic Articulography (EMA)

Ultrasound Imaging

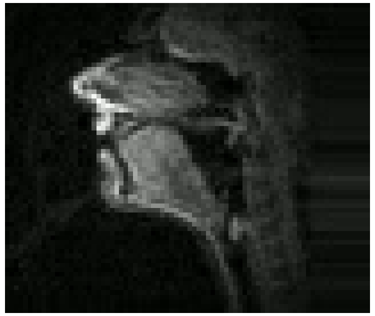
Real-time Magnetic Resonance Imaging (rt-MRI)

Computed Tomography (CT)



Lacks complete view
Disrupts speech

All the articulators are not visible in a single modality



Easy to observe Articulators directly

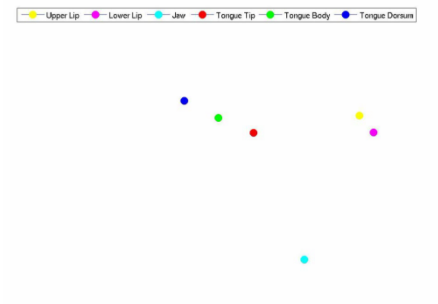
[5] Erik Bresch, Yoon-Chul Kim, Krishna Nayak, Dani Byrd, and Shrikanth Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," IEEE Signal Processing Magazine, vol. 25, no. 3, 2008.
[6] Thomas Hueber, "Ultraspeech-player'' available at http://ultraspeech.com/web/index.php?page=gallery.

Data acquisition:

Electromagnetic Articulography (EMA)

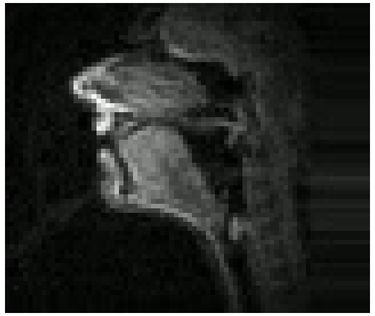
Ultrasound Imaging

Real-time Magnetic Resonance Imaging (rt-MRI)



Lacks complete view
Disrupts speech

All the articulators are not visible in a single modality

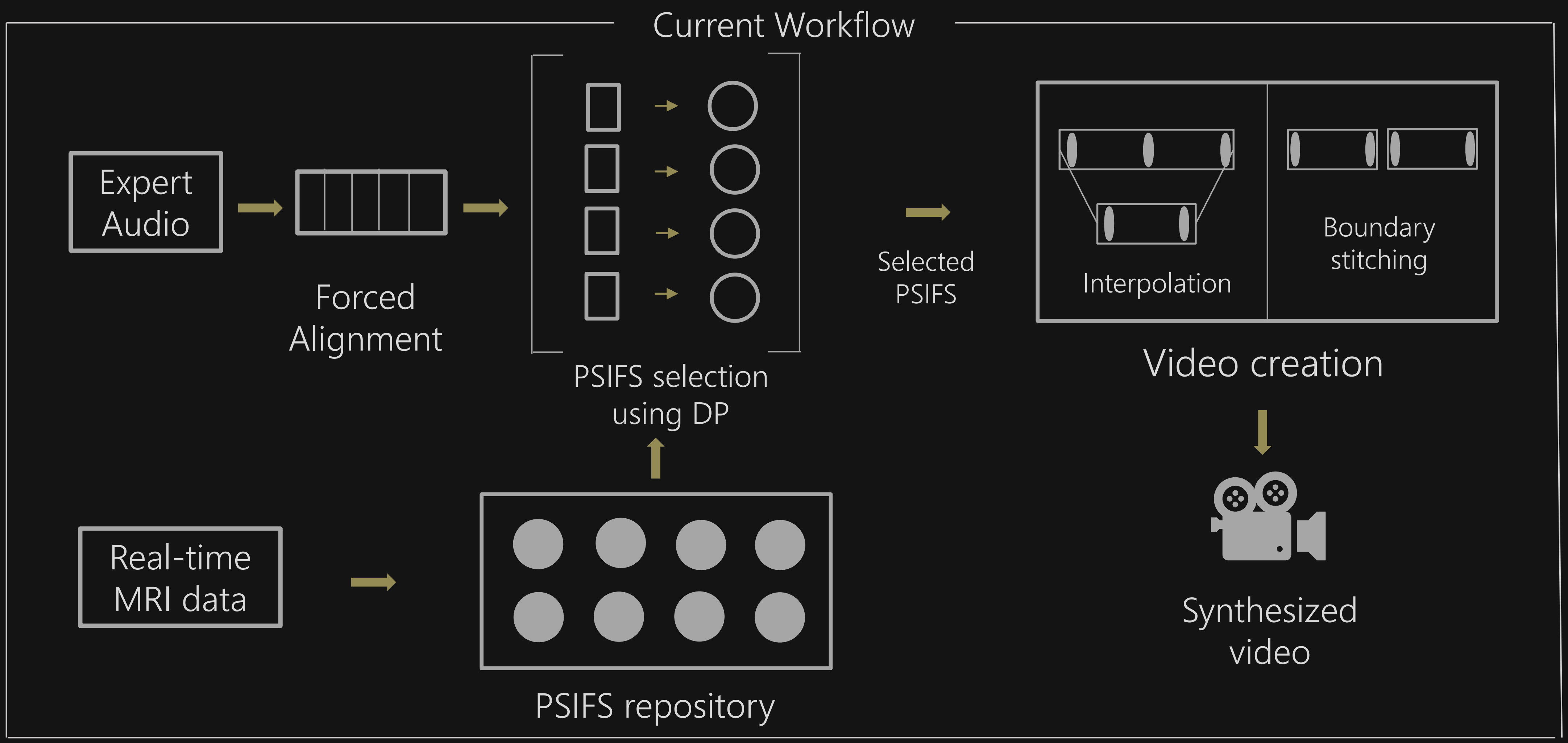


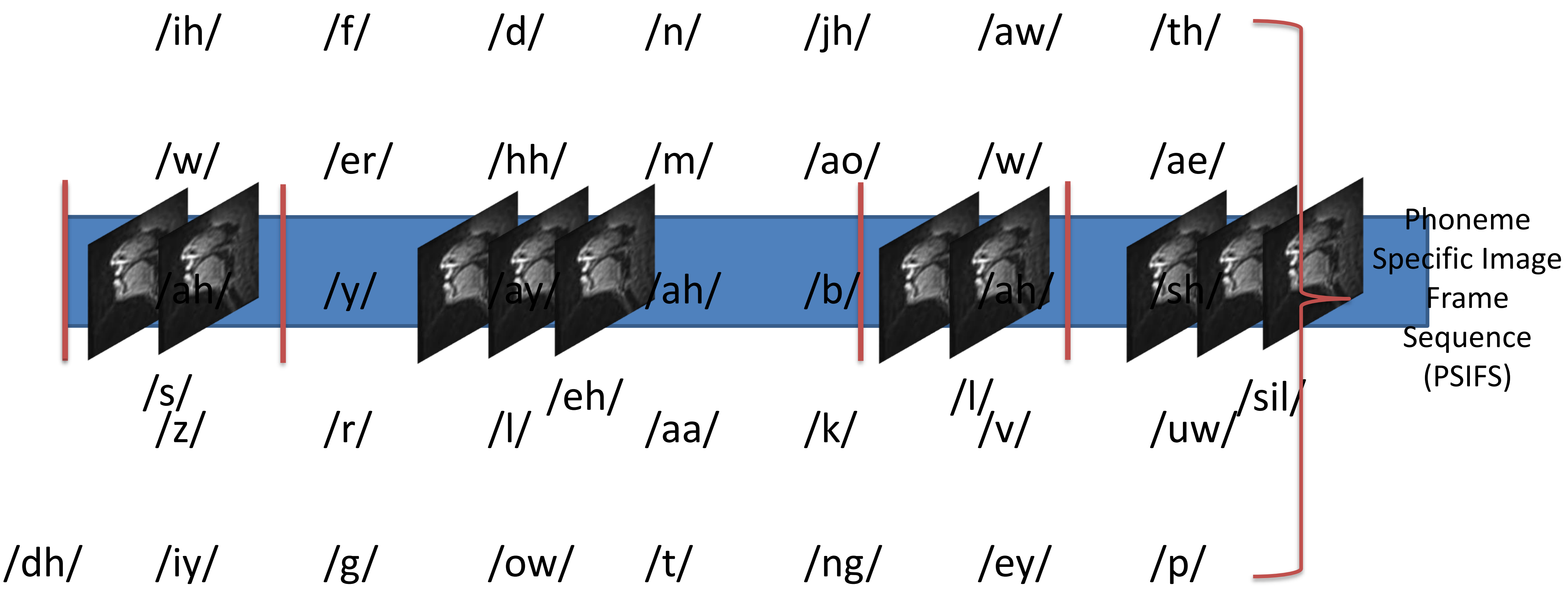
Easy to observe Articulators directly

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.

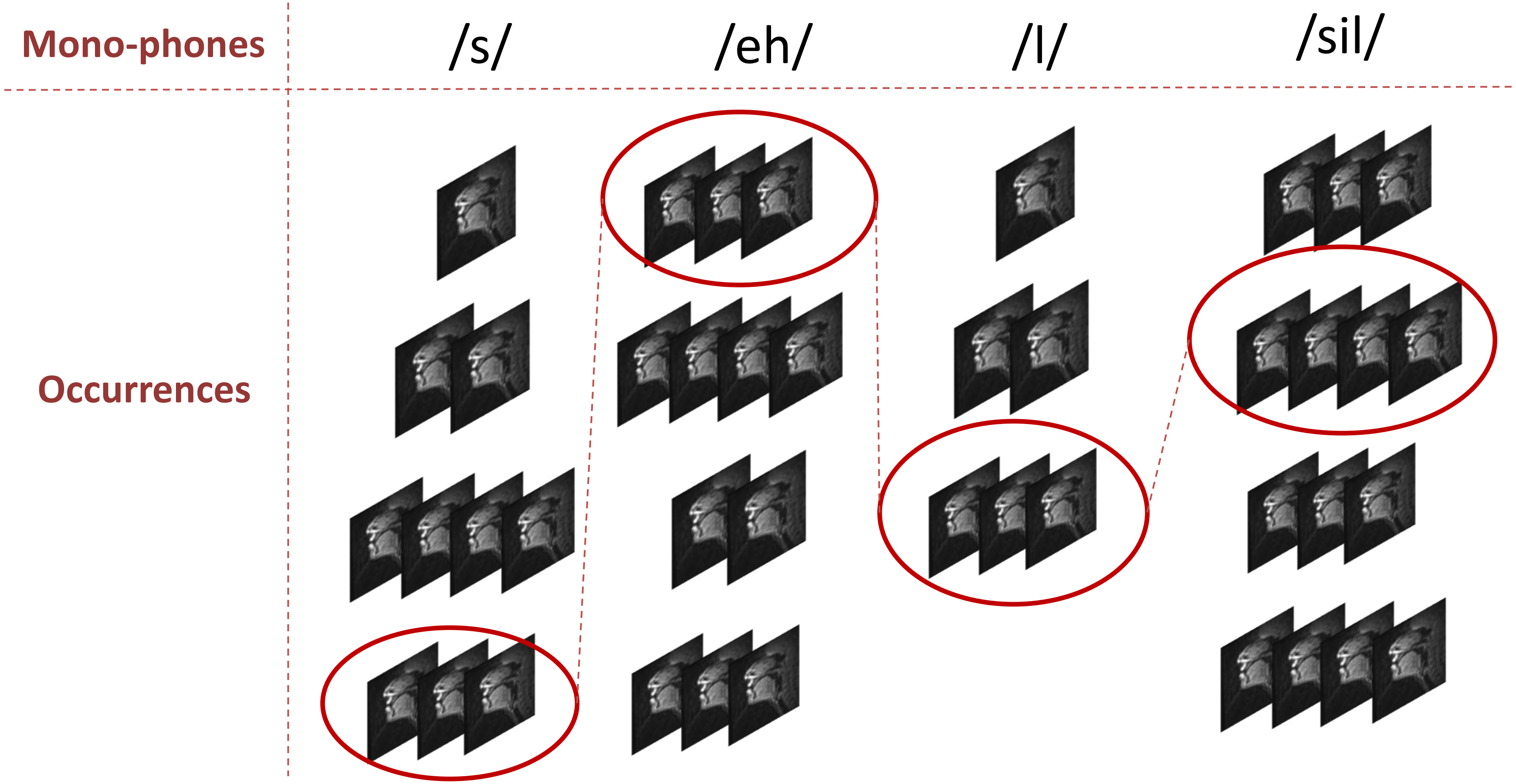
[5] Erik Bresch, Yoon-Chul Kim, Krishna Nayak, Dani Byrd, and Shrikanth Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," IEEE Signal Processing Magazine, vol. 25, no. 3, 2008.
[6] Thomas Hueber, "Ultraspeech-player'' available at http://ultraspeech.com/web/index.php?page=gallery.


Current Workflow
Expert Audio
Forced Alignment
PSIFS selection using DP
Selected PSIFS
Interpolation
Boundary stitching
Video creation
Real-time MRI data
PSIFS repository
Synthesized video



PSIFS repository

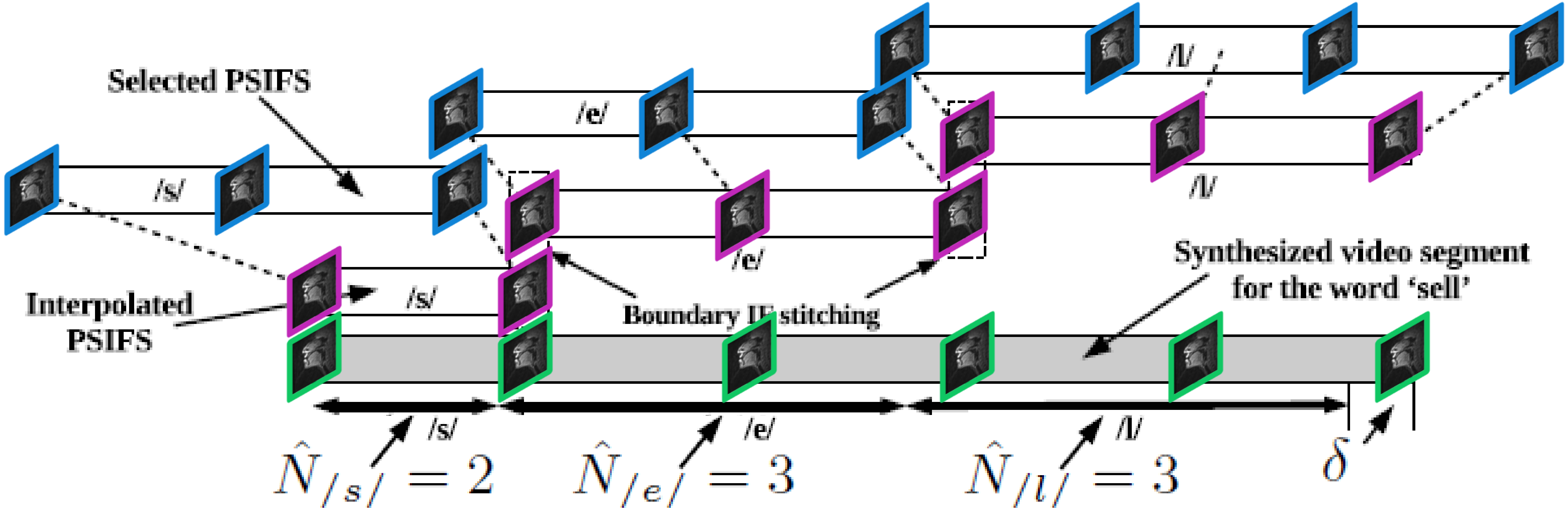
- Select such that the Frobenious norm between the last IF of the previous phoneme and the first IF of the next phoneme should be minimum.



Dynamic Programming

[7] John Watrous, "Theory of quantum information," University of Waterloo Fall, vol. 128, 2011.

- 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

[8] Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, et al., "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research," The Journal of the Acoustical Society of America, vol. 136, no.3, pp. 1307–1311, 2014.

Database

~~#1: • • • taxicab broke down • • •~~

#2: • • • was easy for • • •

⋮

~~#K: • • • with understanding alleviates • • •~~

⋮

#N: • • • worry over silly • • •

⋮

~~#460: • • • the state of • • •~~

Data for the PSIFS repository

taxicab broke down

with understanding alleviates

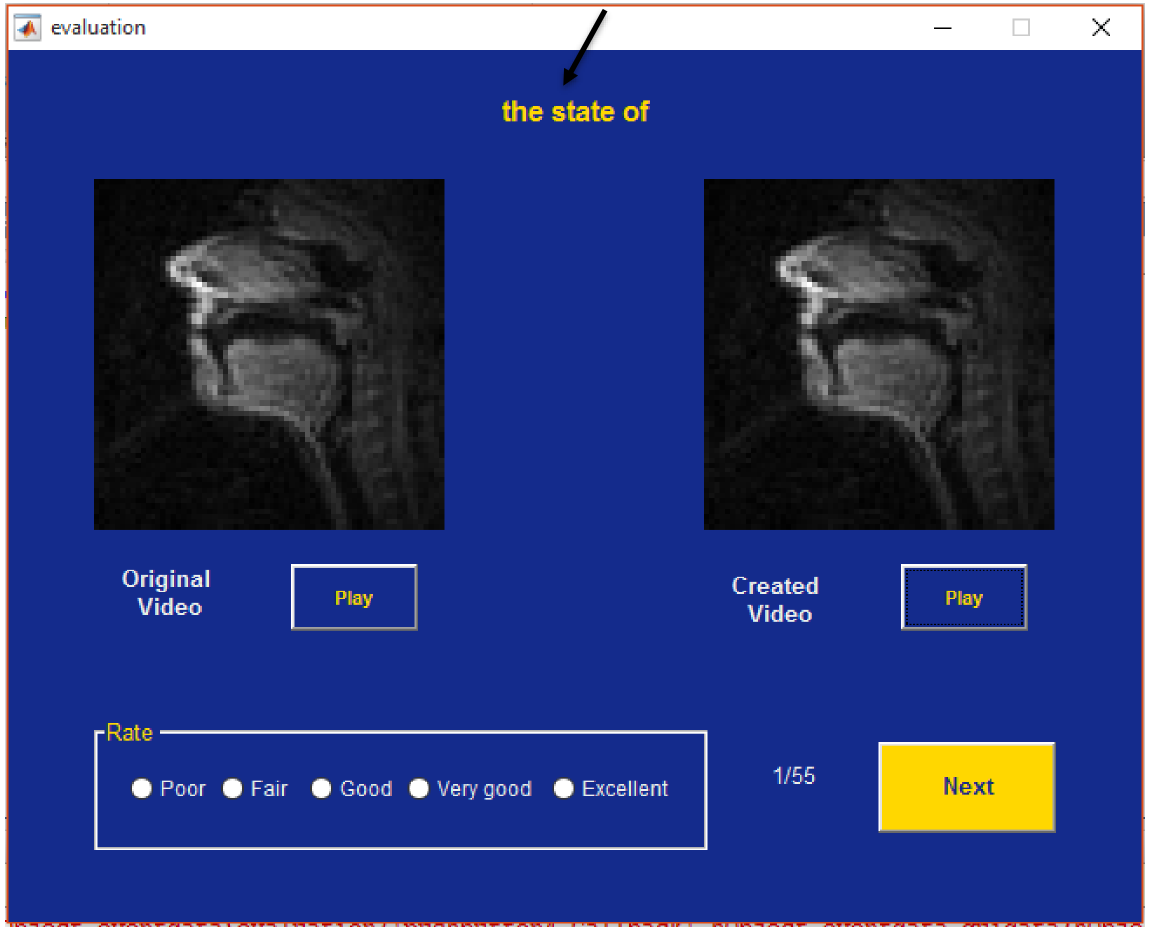
⋮

the state of

Test Stimuli

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.

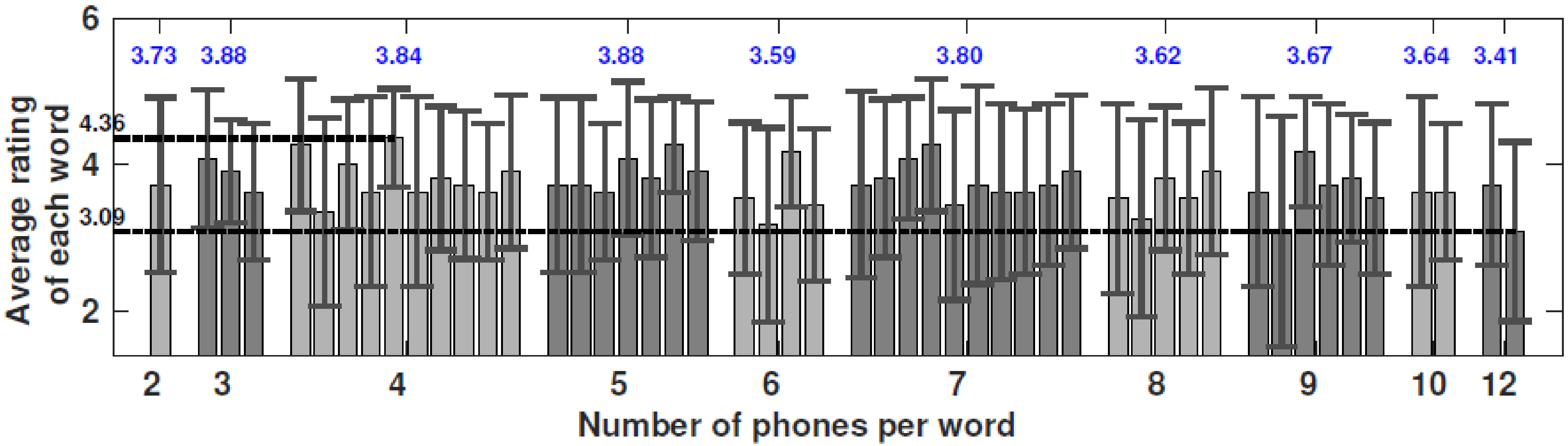
Original **Synthesized**

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

Original **Synthesized**

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