# CONCATENATIVE ARTICULATORY VIDEO SYNTHESIS USING REAL-TIME MRI DATA FOR SPOKEN LANGUAGE TRAINING

*Urvish Desai*[1] *Chiranjeevi Yarra*[2] *Prasanta Kumar Ghosh*[2]

[1]Applied Mathematics, Indian Institute of Technology (ISM), Dhanbad-826004, India
[2]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
urvish.desai2000@gmail.com , {chiranjeeviy, prasantg}@iisc.ac.in

## ABSTRACT

Spoken language training benefits from showing a video of native speakers' articulatory movements to train the second language learners. Typically, the articulatory video is prepared in conjunction with the audio which is collected simultaneously with the articulatory recording. Articulatory video recording requires specialized equipment and, hence, is expensive and time consuming. In this work, we propose a concatenative synthesis approach to obtain articulatory videos for an audio, which may not have a simultaneous articulatory recording. In the training stage of the proposed approach, we make a repository for phoneme specific articulatory image sequence from the available articulatory video. During testing, image sequences are selected from this repository to ensure a smooth transition across phonetic events. The selected image sequences are finally stitched to synthesize the articulatory video for the test audio. Articulatory videos are synthesized for 50 words randomly selected from the MRI-TIMIT database, not seen in the training data. Subjective evaluation on the quality of the synthesized videos using twelve subjects suggests that the videos are close to the original ones with a rating of 3.78 out of 5, where a score of 5 (1) indicates that there is no (great) difference in quality between the original and the synthesized videos.

***Index Terms—*** Articulatory video synthesis, spoken language training, concatenative synthesis, real-time MRI videos

## 1. INTRODUCTION

The pronunciation of the second language (L2) learners, especially spoken English learners, is often effected by several factors [1] [2] [3] that are influenced by their nativity. This happens mainly because the articulatory movements during speaking English are dominated by the articulatory constraints from the speaker's native language [4]. For example, the Tamil people articulate '/t̪/' in place of '/t̪h/', since they do not have aspiration in their native language sounds [4]. It is known that an incorrect phoneme articulation would result in miscommunication [5] [6]. Thus, in L2 training, for example, while learning English, L2 learners need to overcome the influence of their native articulatory gestures in order to have correct articulation while speaking English. It is also known that the L2 learners benefit from a video that shows correct articulation [5] [6]. Such video based feedback are often useful in the applications like computer assisted language learning (CALL). There are a number of reported results, that have shown that the visualization of the correct (from native speakers, we refer to as experts) articulatory movements helps in the pronunciation training [5] [6] [7] [8] [9] [10]. In most of the cases, experts' articulatory movements are captured using real-time motion capture techniques simultaneously with their audio [11] [6] [12] [13]. Further, the articulatory movements, referred to as articulatory video, are added with an augmented reality along with experts' audio to obtain a final video, referred to as augmented articulatory video (AA-video), for the training [8] [6] [14] [15] [16].

Pierre et al. have used the data from electro-magnetic articulography (EMA) to construct an AA-video [7]. In addition to EMA data, they have also used one or more combinations of data from computed tomography (CT), ultrasound imaging and magnetic resonance imaging (MRI) to obtain a better augmented reality in constructing the AA-videos [8]. Similarly, for creating the AA-videos, Engwell et al. have used the combined data from ultrasound imaging and EMA [10]. In a few works, the AA-video is created directly with the articulatory video from ultrasound imaging technique and with an augmented reality [9] [16] [14]. Similarly, Bernd et al. have constructed the AA-video with the articulatory video from MRI with an augmented reality [17]. In constructing the AA-videos, most of the existing works have used an expert from whom both audio and articulatory motion have been recorded. Hence, these techniques have a limitation in using an arbitrary experts' audio from whom direct articulatory measurement is not available. In addition, the data acquisition methods used in all of these techniques require specialized equipment, which is time consuming and expensive [18]. Hence, it becomes challenging to collect articulatory data for a large set of stimuli from multiple experts.

Various articulatory data acquisition methods have their own advantages and disadvantages [18]. The CT method has good temporal and spatial resolution and captures pharyngeal structures. Hence, it is the most suitable modality for creating AA-videos. However, its main disadvantage is that it exposes the subject to radiation. Similarly, EMA has a high temporal resolution but it only tracks a few sensors placed on the articulators, thus lacking a complete view of the vocal tract. It cannot also capture pharyngeal structures. In the absence of a complete mid-sagittal view, EMA based video may not be effective for L2 learners. In addition, Meenakshi et al. have shown that due to the presence of EMA sensors in the vocal tract, the audio during EMA recording differs significantly from the natural voice of the subject [19]. To circumvent these limitations, most of the works have used ultrasound and rt-MRI method, which are non-invasive and safe [18]. The ultrasound imaging has high temporal resolution and detects only first air-tissue boundary and, hence, is not suitable for anterior tongue tip and lip imaging. On the other hand, the rt-MRI captures pharyngeal structures and, hence, reduces the effort in augmented reality. Even without augmented reality, it is easy to notice the articulators in an rt-MRI video. Thus, there is flexibility in using rt-MRI modality directly for L2 training. However, rt-MRI recording setup is expensive and also provides a relatively poor spatial and temporal resolution [13]. But, depending on the frame rate, it could be suitable to create an AA-video since a minimum frame rate of 15 frames per second is good enough for such purpose [20].

In this work, we propose an automatic articulatory video synthesis method corresponding to an arbitrary expert's audio even though direct articulatory measurements may not be available for the expert. It should be noted that, in the proposed method, audio is not synthesized rather taken from an expert and the corresponding image sequence for the video is synthesized. Similar to this problem, audiovisual synthesis approaches have been addressed in the literature for synthesizing the movements of visible articulators (when look-

ing at a speaker's face) such as the lips and the cheeks [21]. However, synthesis of other articulatory movements such as the tongue and the velum is challenging and has been less explored [22] [23]. This is due to the challenges involved in the collection of such articulatory data [22]. In the audio visual synthesis, the video synthesis approaches are typically inspired by speech synthesis techniques among which the popular ones are – concatenative and statistical parametric approaches [21]. Among these two, concatenative synthesis is commercially used in many systems [22]. In addition, the quality of the synthesized signal obtained from the statistical approaches depends on the training data size, typically, requiring reasonably large amount of data [24]. In general, the articulatory data corpora are less in size [13], hence, we, in this work, use concatenative synthesis approach for the articulatory video synthesis.

Given an audio, in order to synthesize a corresponding video, we obtain phonetic boundaries in the audio using forced-alignment. Further, we find the best representative image frames (IFs) for each phoneme in a given context as well as maintain smoothness across video frames. This, in turn, demands that the available corpus be rich with phoneme in different contexts. For this purpose, we use MRI-TIMIT [13], which consists of rt-MRI videos of subjects speaking phonetically balanced sentences. Finally, we interpolate the selected IFs to synchronize with the audio and concatenate the interpolated IFs by stitching operation at the boundary. We evaluate the synthesized video quality subjectively using a set of 12 raters and 50 words randomly from the MRI-TIMIT data. The average quality rating is found to be 3.78 out of 5 when the raters rate the synthesized video quality with respect to the corresponding original video.

## 2. REAL-TIME MRI (rt-MRI) DATABASE

MRI-TIMIT is a phonetically rich database comprising rt-MRI videos, i.e., rt-MRI data with synchronized audio. The rt-MRI data is primarily an IF sequence of the mid-sagittal view of a speaker speaking an utterance. The rt-MRI data was captured at a frame rate of 23.18 frames per second with an image resolution of 68×68 pixels in gray scale. The audio was simultaneously recorded with rt-MRI data at a sampling frequency of 20kHz inside an MRI scanner using a fiber-optic microphone. The data was collected from two male and two female speakers of American English speaking 460 TIMIT sentences. Along with the videos, text stimuli are available for all utterances. Among the four speakers, we consider data from one female speaker for our experiments and extract audio from the rt-MRI video of each utterance using FFmpeg. We estimate phonetic transcriptions and its aligned boundaries using forced-alignment implemented using the Kaldi tool kit [25] considering deep neural network based acoustic models (Karels' implementation) learnt from the fisher English [26] data. In the forced-alignment, we use a lexicon obtained by combining the CMU [27] and TIMIT [28] pronunciation dictionaries. From the data, it is also observed that the total number of unique phones is 40.

## 3. PROPOSED APPROACH

Block diagram in Figure 1 shows the three steps involved in the proposed video synthesis. In the first step, we construct a phoneme specific IF sequences (PSIFS) repository $\mathscr{V}$ using a training articulatory data for the set ($\mathscr{Q}$) of all the phonemes in a given context, where $\mathscr{Q} = \{Q(1), Q(2), \ldots, Q(M)\}$ is a collection of all the context dependent phonemes, whose total count is $M$, in the training set in which $Q(i)$ is $i$-th context dependent phoneme. $\mathscr{V} = \{\mathscr{V}_{Q(1)}, \mathscr{V}_{Q(2)}, \ldots, \mathscr{V}_{Q(M)}\}$ in which $\mathscr{V}_{Q(i)}$ is a collection of $\left|\mathscr{V}_{Q(i)}\right|$ PSIFS, where $\mathscr{V}^k_{Q(i)} = \{\mathscr{V}^k_{Q(i)}(l),\ 1 \leq l \leq N^k_{Q(i)}\}$ denotes the $k$-th PSIFS belonging to $Q(i)$ and $N^k_{Q(i)} = \left|\mathscr{V}^k_{Q(i)}\right|$, the cardinality of $\mathscr{V}^k_{Q(i)}$. $\mathscr{V}^k_{Q(i)}(l)$ is the $l$-th image frame of 68×68 pixels in the $k$-th PSIFS for $Q(i)$. In the second step, given a test audio and its text, we apply forced-alignment to obtain phonetic boundaries. From this, we obtain a context dependent phoneme set $\mathscr{P} = \{P(1), P(2), \ldots, P(N)\}$ and the corresponding durations $\{d_{P(1)}, d_{P(2)}, \ldots, d_{P(N)}\}$ for each $P(i)$, where $N$ is the total number of forced-aligned phonemes in the test audio and $\mathscr{P} \subseteq \mathscr{Q}$. Using $\mathscr{P}$ and $\mathscr{V}$, we find the best PSIFS $\hat{V}_{P(i)}$ for each $P(i)$ using a dynamic programming (DP) approach by ensuring maximum smoothness across the selected PSIFS. In the third step, we interpolate the selected PSIFS $\hat{V}_{P(i)}$ to synchronize with its corresponding duration $d_{P(i)}$. Following this, we perform stitching between two boundary IFs of the PSIFS of every two consecutive phonemes. Finally, we combine the audio to obtain a synthesized video.

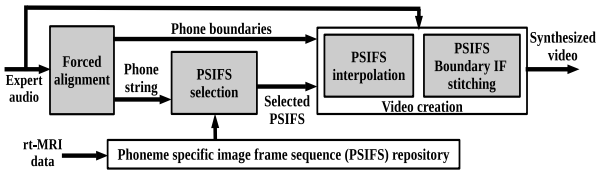


**Fig. 1**. Block diagram illustrating the steps involved in the proposed approach for video synthesis.

### 3.1. Phoneme specific image frame sequences (PSIFS) repository

Given time-aligned phoneme transcriptions of an utterance and its respective articulatory data in the training set, we obtain context dependent phonemes and corresponding durations. Following this, we compute start and end IF locations in the articulatory data for every context dependent PSIFS. Let $x$ be a start or end time of a context dependent phoneme, then its respective start or end IF index is computed as $\lceil x \times \mathscr{F} \rceil$, where $\mathscr{F}$ is the video frame rate and $\lceil x \rceil$ is the lowest integer higher than $x$. Figure 2 illustrates the computation of IF indices for an exemplary word 'sell' from the rt-MRI data. The word has three phonemes '/s/, /e/, /l/'. As the end time of a phoneme is identical to the start time of next phoneme, from the figure, it is observed that the start IF indices of PSIFS for the phonemes '/e/' and '/l/' are identical to the end IF indices of PSIFS for the phonemes '/s/' and '/e/' respectively.

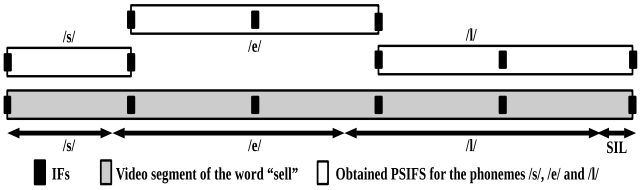


**Fig. 2**. An example word 'sell' illustrating the preparation of PSIFS repository.

### 3.2. PSIFS selection

Given an expert audio's $\mathscr{P} = \{P(1), P(2), \ldots, P(N)\}$, we find the best PSIFS $\hat{V}_{P(i)}$ for each $P(i)$ to ensure smoothness across the selected PSIFS. Since the PSIFS are obtained from the original naturally recorded rtMRI video, we assume that the smoothness within the selected PSIFS is automatically ensured. Thus, we use a cost function that minimizes the discontinuities at the boundaries of the selected PSIFS given by,

$$\left\{\hat{V}_{P(i)},\ _{1\leq i\leq N}\right\} = \underset{\substack{V_{P(i)}\in\mathscr{V}_{P(i)}\\ \forall i\in\{1:N\}}}{\arg\min} \sum_i \mathscr{D}\left(V_{P(i)}(N_{P(i)}), V_{P(i+1)}(1)\right) \quad (1)$$

where, $\mathscr{D}(A, B)$ is the Frobenius norm [29] between two IFs A and B. We assume that the cost function $\mathscr{D}$ ensures smoothness across frames in the synthesized video. The optimization problem is solved using DP. The detailed steps for solving (1) are provided in Algorithm 1.

**Algorithm 1** PSIFS selection using DP. Input: $\mathscr{P} = \{P(1), P(2), \ldots, P(N)\}$, $\mathscr{V} = \{\mathscr{V}_{P(1)}, \mathscr{V}_{P(2)}, \ldots, \mathscr{V}_{P(N)}\}$. Output: $\hat{V}_{P(i)}, 1 \leq i \leq N$

1: Initialization: $T_{P(i)} = \left|\mathscr{V}_{P(i)}\right|; C_1(r) = 0 \ \forall \ r \in \{1 : T_{P(1)}\}$
2: **for** each phone index $i$ from 2 to $N$ **do**
$\forall \ r \in \{1 : T_{P(i)}\}$
$C_i(r) = \min_{j \in \{1:T_{P(i-1)}\}} \left\{C_{i-1}(j) + \mathscr{D}\left(\mathscr{V}^j_{P(i-1)}(N_{P(i-1)}), \mathscr{V}^r_{P(i)}(1)\right)\right\}$
$k_i(r) = \arg\min_{j \in \{1:T_{P(i-1)}\}} \left\{C_{i-1}(j) + \mathscr{D}\left(\mathscr{V}^j_{P(i-1)}(N_{P(i-1)}), \mathscr{V}^r_{P(i)}(1)\right)\right\}$
3: **end for**
4: Back tracking: $\eta_N = \arg\min_{r \in \{1:T_{P(N)}\}} \{C_N(r)\}, \hat{V}_{P(N)} = \mathscr{V}^{\eta_N}_{P(N)}$
5: **for** each frame $i$ from $N-1$ to 1 **do**
$\eta_i = k_{i+1}(\eta_{i+1}), \hat{V}_{P(i)} = \mathscr{V}^{\eta_i}_{P(i)}$
6: **end for**

### 3.3. Video creation

#### *3.3.1. Interpolation*

The duration of the selected PSIFS for $P(i)$ may not match with the duration of the test expert's audio, i.e., $d_{P(i)}$. Hence, we propose an interpolation technique to alter the length of $\hat{V}_{P(i)}$ accordingly. The interpolation is done in two steps. In the first step, the required number of frames $\hat{N}_{P(i)}$ belonging to $P(i)$ is computed as

$$\hat{N}_{P(i)} = \left\lceil \left(\sum_{j=1}^{i} d_{P(j)}\right) \times \mathscr{F} - \sum_{j=1}^{i-1} (\hat{N}_{P(j)} - 1) \right\rceil \qquad (2)$$

In (2), $\left(\sum_{j=1}^{i} d_{P(j)}\right)$ is the total time at the end of $P(i)$ and $\sum_{j=1}^{i-1}(\hat{N}_{P(j)} - 1)$ is the total number of frames constructed in synthesized video before $P(i)$. The $-1$ in (2) is due to merging of two boundary (discussed in the next section) IFs of PSIFS to one frame at the boundary of every $P(i)$.

Equation (2) has an advantage compared with an alternative computation of $\hat{N}_{P(i)}$, directly from $d_{P(i)}$ as $\hat{N}_{P(i)} = \lceil d_{P(i)} \times \mathscr{F} \rceil$ unlike the cumulative sum in (2). It is easy to show that, in an utterance, the total rounding error in (2) due to the $\lceil . \rceil$ operation is lesser than that in the above mentioned alternative computation. This is because, in the alternative computation, the error occurs at the end of each phoneme and it is accumulated $N$ times at the end of the utterance. However, in (2), the error occurs only at the end of the utterance and it is $\left|\left(\sum_{i=1}^{N}(\hat{N}_{P(i)} - 1)\right) \times \mathscr{F} - \sum_{i=1}^{N} d_{P(i)}\right|$, referred to as $\delta$. To compensate for this error, we add silence of $\delta$ duration at end of the test experts audio. This, in turn, removes the mismatch between the duration of the audio and that of the synthesized video.

In the second step, we convert $N_{P(i)}$ values at each pixel location in the selected PSIFS $\hat{V}_{P(i)}$ to $\hat{N}_{P(i)}$ values using linear interpolation technique. Hence, we obtain of total of $\hat{N}_{P(i)}$ frames in the interpolated PSIFS $\hat{V}^I_{P(i)}$. Figure 3 shows $\hat{V}_{P(i)}$ and $\hat{V}^I_{P(i)}$ of the phonemes '/s/,/e/,/l/' in the word 'sell' as chosen in Figure 2. From the figure, it is observed that the total number of IFs in $\hat{V}_{P(i)}$ are 3, 3 and 4 for '/s/,/e/,/l/' phonemes and their respective $\hat{N}_{p(i)}$ are 2, 3 and 3. It should be noted that, after interpolation, the start and end IFs in the $\hat{V}^I_{P(i)}$ are kept identical to those in the $\hat{V}_{P(i)}$.

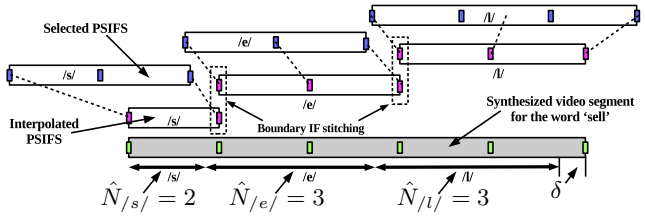


**Fig. 3**. An illustrative example explaining the PSIFS interpolation and video stitching for the word 'sell'.

#### *3.3.2. PSIFS boundary IF stitching*

We assume that the first and last IFs in the selected PSIFS for a phoneme corresponds to the transitions to the neighboring phonemes. In this work, we consider the IF corresponding to a inter-phoneme transitions from two boundary IFs of two consecutive phonemes and stitch them using function $f$. For this, in $\hat{V}^I_{P(i)}$, we consider the start and end IFs as representative of the transition from $P(i-1)$ to $P(i)$ and from $P(i)$ to $P(i+1)$ respectively. For example, in Figure 3, the start IF of $\hat{V}^I_{/e/}$ and the end IF of $\hat{V}^I_{/s/}$ are considered to represent the same inter-phoneme transition i.e., from '/s/' to '/e/' and on those IFs, the stitching function $f$ is applied. We choose the function $f$ as the average of two corresponding pixels in these two IFs. We hypothesize that the average would result in smooth phoneme transition in the synthesized video.

## 4. EXPERIMENTS AND RESULTS

### 4.1. Experimental setup

The rt-MRI videos of 460 sentences from one of the female subjects is used for the experiments. For the evaluation, we synthesize the videos for a random set of 50 words, which occur only once in the entire corpora. We consider the audio in the rt-MRI videos of the test words as the test expert's audio for the synthesis. While the expert's audio can be taken from any other subject, in this work, those are chosen from the rt-MRI videos. This is done to ensure that during evaluation, the ground truth rt-MRI video is available for each test audio. We perform the synthesis under mono-phone context, i.e., $M$ is the total number phonemes in the corpus. In creating the PSIFS repository, we consider all the utterances in the corpora excluding the utterances containing those 50 words used for evaluation.


**Fig. 4**. Graphical user interface (GUI) used in the subjective evaluation.

As the mid-sagittal view of the subject does not have the same orientation in all videos, we perform required image translation and rotation to ensure that the mid-sagittal view in all video frames have

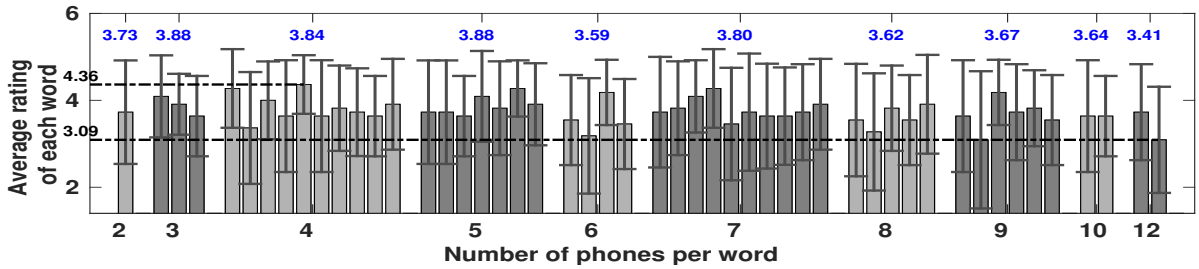


**Fig. 5**. Average ratings obtained across all the evaluators for each word; the blue colored numbers indicate the average rating across all the evaluators and all the words for a given number of phonemes in a word.

identical orientation. For this, we randomly select an image frame as a reference image and rotate the remaining image frames to have the same orientation as in the reference image. Following this, we compute an angle of rotation and translation vector based on the lines joining the bottom nose point and the lower jaw point considering the nose point on the reference image as the reference point.

### 4.2. Subjective evaluation

We conduct the subjective evaluation using a set of 12 evaluators (6 males and 6 females). The evaluators are in the age group of 20 to 33 years with an average age of 23.58 years ($\pm$3.60). The evaluators are undergraduate and graduate engineering students. None of the evaluators has any vision problems. All the evaluators can read, write and speak English fluently.

#### *4.2.1. Description of the evaluation set-up*

In the evaluation, we present original and synthesized videos for every word to the evaluator in the context of previous and next words. For this, in the original video containing previous, target and next words, we replace the target word's video segment with its synthesized video. Hence, it helps in quantifying the quality of the synthesized video at the word boundaries. The average duration of the videos used for evaluation is found to be 1.29 seconds ($\pm$0.39) and all the evaluators found the duration of the words to be comfortable for evaluation. We ask the evaluator to rate the presented videos using the following five categories:

- 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.

This evaluation is done using a graphical user interface (GUI) developed using MATLAB R2015a as shown in Figure 4. It allows the evaluator to play the original video and the synthesized video separately as many times as he/she wants. The GUI displays the target word transcription including its previous and next words. The GUI provides radio buttons for obtaining the evaluator ratings. The GUI also displays the progress of the evaluation. To know the consistency of the evaluator, we randomly repeat 5 synthesized videos. All the evaluators are found to have more than 60% matching in the ratings of the repeated words.

#### *4.2.2. Results and discussion*

From the evaluator ratings, it is found that the quality of the synthesized videos is 3.78 ($\pm$1.07) when averaged across all the 12 evaluators and all 50 stimuli. This indicates that the quality of the synthesized videos is not significantly different from that of the original video. Figure 5 shows the average rating across the evaluators for each word with respect to the number of phonemes in a word. From the figure, it is observed that the highest average rating is 4.36, which occurs for the word "broke", which has 4 phonemes. Similarly, the least average rating is 3.09 for the words "crisscrossed" and "understanding", which have 9 and 12 phonemes respectively. Figure 6 shows the synthesized video frames corresponding to the word "broke", which has four frames.

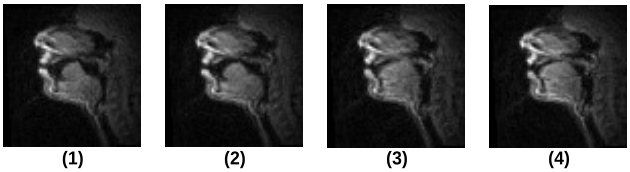


**Fig. 6**. Four synthesized video frames corresponding to the word 'broke'(/b/,/r/,/oʊ/,/k/) which has the highest rating among all 50 words used.

From Figure 5, comparing average ratings for every number of phonemes per word, it is also observed that the ratings do not directly depend on the number of phonemes in a word. However, the number of phonemes averaged across the words with less than and greater than 3.78 (overall average) is found to be 5.7 and 6.8 respectively. This indicates that, on an average, the words with more phonemes have lower quality. This is because the words containing more phonemes have more boundaries to smooth and, hence, could result in more disruptions in the videos. In general, there could be cases where the target phoneme duration is largely different from the selected PSIFS phoneme duration. With large number of phonemes in a word, such large deviations is more likely to occur.

## 5. CONCLUSIONS

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. Further investigations are required to develop better techniques for image stitching as well as for PSIFS selection and interpolation.

## 6. ACKNOWLEDGEMENT

Authors thank all the subjects participated in the evaluation and the Pratiksha Trust for their support.

## 7. REFERENCES


[1] Chiranjeevi Yarra, Om D Deshmukh, and Prasanta Kumar Ghosh, "Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5845–5849, 2017.

[2] Supriya Nagesh, Chiranjeevi Yarra, Om D Deshmukh, and Prasanta Kumar Ghosh, "A robust speech rate estimation based on the activation profile from the selected acoustic unit dictionary," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5400–5404, 2016.

[3] Chiranjeevi Yarra, Om D Deshmukh, and Prasanta Kumar Ghosh, "A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection," *Speech Communication*, vol. 78, pp. 62–71, 2016.

[4] Sailaja Pingali, *Indian English*, Edinburgh University Press, 2009.

[5] Ambra Neri, Catia Cucchiarini, and Helmer Strik, "Feedback in computer assisted pronunciation training: technology push or demand pull?," *International Conference on Spoken Language Processing (ICSLP)*, pp. 1209–1212, 2002.

[6] Maxine Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[7] Pierre Badin, Yuliya Tarabalka, Frédéric Elisei, and Gérard 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.

[8] Pierre Badin, Frédéric Elisei, Gérard Bailly, and Yuliya Tarabalka, "An audiovisual talking head for augmented speech generation: models and animations based on a real speakers articulatory data," *Internation Conference on Articulated Motion and Deformable Objects*, pp. 132–143, 2008.

[9] Pierre Badin, Atef Ben Youssef, Gérard Bailly, Frédéric 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.

[10] Olov Engwall, "Can audio-visual instructions help learners improve their articulation?–an ultrasound study of short term changes.," *Proceedings of Interspeech*, pp. 2631–2634, 2008.

[11] Thomas Hueber, Gérard Chollet, Bruce Denby, and Maureen Stone, "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," *Proceedings of International Seminar on Speech Production (ISSP)*, pp. 365–369, 2008.

[12] Alan A Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," *Workshop on Phonetics and Phonology in ASR, Saarbruecken, Germany*, pp. 1–13, 2000.

[13] 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.

[14] Dominic W Massaro and Joanna Light, "Using visible speech to train perception and production of speech for individuals with hearing loss," *Journal of speech, Language, and hearing research*, vol. 47, no. 2, pp. 304–320, 2004.

[15] Gérard Bailly, Pierre Badin, Denis Beautemps, and Frédéric Elisei, "Speech technologies for augmented communication," *Proceedings of Computer Synthesized Speech Technologies: Tools for Aiding Impairment, Mullennix, J. and Stern, S., Eds.: IGI Global, Medical Information Science Reference*, pp. 116–128, 2010.

[16] Thomas Hueber, "Ultraspeech-player: intuitive visualization of ultrasound articulatory data for speech therapy and pronunciation training.," *Proceedings of Interspeech*, pp. 752–753, 2013.

[17] Bernd J Kröger, Verena Graf-Borttscheller, and Anja Lowit, "Two and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders," *Proceedings of Interspeech*, pp. 2639–2642, 2008.

[18] 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.

[19] Nisha Meenakshi, Chiranjeevi Yarra, BK Yamini, and Prasanta Kumar Ghosh, "Comparison of speech quality with and without sensors in electromagnetic articulograph ag 501 recording," *Proceedings of Interspeech*, pp. 935–939, 2014.

[20] Jessie YC Chen and Jennifer E Thropp, "Review of low frame rate effects on human performance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 37, no. 6, pp. 1063–1076, 2007.

[21] Wesley Mattheyses and Werner Verhelst, "Audiovisual speech synthesis: An overview of the state–of–the–art," *Speech Communication*, vol. 66, pp. 182–217, 2015.

[22] B Theobald, "Audiovisual speech synthesis," *International Congress on Phonetic Sciences*, pp. 285–290, 2007.

[23] Björn Granström, "Towards a virtual language tutor," *InSTIL/ICALL Symposium*, 2004.

[24] Tomoki Koriyama and Takao Kobayashi, "A comparison of speech synthesis systems based on gpr, hmm, and dnn with a small amount of training data," *Proceedings of Interspeech*, 2015.

[25] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," *IEEE workshop on automatic speech recognition and understanding (ASRU)*, 2011.

[26] Christopher Cieri, David Miller, and Kevin Walker, "The Fisher Corpus: a resource for the next generations of speech-to-text.," *International conference on Language Resources Evaluation*, vol. 4, pp. 69–71, 2004.

[27] "CMU pronouncing dictionary – version 0.7," *Available from: http://www.speech.cs.cmu.edu/cgi-bin/cmudict, last accessed on 26-10-2017*.

[28] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

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