Interspeech 2018
2-6 September 2018, Hyderabad

Automatic visual augmentation for concatenation based synthesized articulatory videos from real-time MRI data for spoken language training

Chandana S, Chiranjeevi Yarra[1], Ritu Aggarwal, Sanjeev Kumar Mittal, Kausthubha N K, Raseena K T, Astha Singh, Prasanta Kumar Ghosh[2]

Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

{[1]chiranjeeviy,[2]prasantg}@iisc.ac.in

Abstract

For the benefit of spoken language training, concatenation based articulatory video synthesis has been proposed in the past to overcome the limitation in the articulatory data recording. For this, real time magnetic resonance imaging (rt-MRI) video image-frames (IFs) containing articulatory movements have been used. These IFs require a visual augmentation for better understanding. We, in this work, propose an augmentation method using pixel intensities in the regions enclosed by the articulatory boundaries obtained from air-tissue boundaries (ATBs). Since, the pixel intensities reflect the muscle movements in the articulators, the augmented IFs could provide realistic articulatory movements, when we color them accordingly. However, the ATB manual annotation is time consuming; hence, we propose to synthesize ATBs using the ATBs from a few selected frames that have been used in synthesizing the articulatory videos. We augment a set of synthesized articulatory videos for 50 words obtained from the MRI-TIMIT database. Subjective evaluation on the quality of the augmented videos using twenty-one subjects suggests that the videos are visually more appealing than the respective synthesized rt-MRI videos with a rating of 3.75 out of 5, where a score of 5 (1) indicates that the augmented video quality is excellent (poor).

1. Introduction

The pronunciation of the second language (L2) learners, especially learning English, is often effected by several factors [1–3] that are influenced by their nativity. This happens mainly because the articulatory movements while speaking English are dominated by the articulatory constraints from the speaker's native language [4]. In order to overcome these constraints, a video that shows correct articulation is used as a feedback to the L2 learners in the applications like computer assisted language learning (CALL). There have been several results that shows the visualization of the correct (from native speakers, referred as experts) articulatory movements which helps in the pronunciation training [5–10]. In most of the cases, for the training, experts' articulatory movements are captured using real-time motion capture techniques simultaneously with their audio [6, 11–13]. 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 as augmented articulatory video (AA-video) [6, 8, 14–16].

In the existing works, the AA-videos have been constructed using one or more combinations of the articulatory data from electro-magnetic articulography (EMA), computed tomography (CT), ultrasound imaging and real time magnetic resonance imaging (rt-MRI) [7–10, 14, 16, 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 expert's 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]. However, in the recent past, Desai et al. have proposed a concatenation based synthesis approach to obtain an articulatory video for an expert audio which does not have simultaneous articulatory recordings [19]. In their work, they have used rt-MRI videos containing image frames (IFs) of pharyngeal structures in gray scale. We observe that the articulators constituted in those structures do not have a realistic view; hence, the synthesized videos are less self explanatory to the L2 learners. However, we hypothesize that an augmented reality can be added automatically to those videos. Thus, an AA-video can be obtained for an audio of an expert for whom direct articulatory measurement is not available.

In this work, we add augmented reality to the articulators in each IF belonging to the synthesized articulatory videos. For this, we propose to use pixel values in the IF regions enclosed by the air-tissue boundaries (ATBs, blue and green colored contour shown in Figure 1b) that constitute the articulators [20]. Instead of using ATBs of all the IFs in the synthesized videos, we consider the ATBs of few IFs from a repository which are used in a concatenation based approach [19]. This results in a less number of IFs for the ATB annotation thereby requiring less time. In order to obtain ATBs for all the IFs, we propose an ATB synthesis approach in line with the concatenation based articulatory video synthesis approach. Further, using these ATBs, we apply a knowledge based coloring approach to those structures, for which we propose a set of rules. We evaluate the AA-video quality subjectively using a set of 21 evaluators and 50 words randomly chosen from the MRI-TIMIT data [21]. The average quality rating is found to be 3.75 out of 5 when the evaluators rate the AA-video quality with respect to the corresponding synthesized articulatory video.

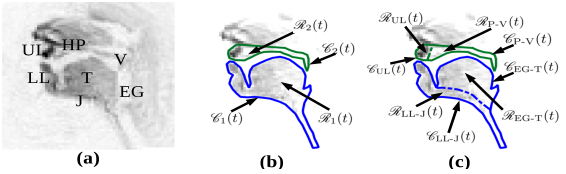

Figure 1: *Exemplary rt-MRI IF indicating a) anatomical regions b) ATBs $\mathscr{C}_1(t)$ and $\mathscr{C}_2(t)$ and respective enclosed regions $\mathscr{R}_1(t)$ and $\mathscr{R}_2(t)$. c) sub-regions $\mathscr{R}_{\text{LL-J}}(t)$, $\mathscr{R}_{\text{EG-T}}(t)$, $\mathscr{R}_{\text{UL}}(t)$ and $\mathscr{R}_{\text{P-V}}(t)$ and the respective boundaries $\mathscr{C}_{\text{LL-J}}(t)$, $\mathscr{C}_{\text{EG-T}}(t)$, $\mathscr{C}_{\text{UL}}(t)$ and $\mathscr{C}_{\text{P-V}}(t)$*

2. Database

MRI-TIMIT [21] 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 (contains pharyngeal structures) 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 data was collected from two male and two female speakers of American English speaking 460 TIMIT sen-

Authors thank Pratiksha Trust for their support.

10.21437/Interspeech.2018-1570

tences. Among these 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 [22]. Following the work proposed by Pattem et al. [20], on each IF in the IF sequence, we annotate two ATBs $\mathscr{C}_1(t)$ & $\mathscr{C}_2(t)$ that pass through different anatomical regions in the mid-sagittal plane, namely, 1) Upper lip (UL), hard palatte (HP) and Velum (V) 2) Jaw (J), Lower lip (LL), Tongue (T) and Epiglottis (EG) as shown in Figure 1a.

3. Background and motivation

In a concatenation based video synthesis using rt-MRI data, the utterance of an expert audio is represented as a sequence of smaller acoustic units (AUs) which are, in general, context dependent phonemes [19]. For each AU, an IF sequence is selected from a repository containing many IF sequences and its length is interpolated according to the AU duration. Further, all the selected IF sequences of the AUs are stitched together sequentially and are combined with the expert audio to obtain a synthesized video. In order to ensure a smooth transition at AU boundaries, two boundary IFs of two consecutive AUs are merged into one IF to represent the boundary between those AUs.

In the augmented reality, often, the muscle movements in the soft tissues have been considered to obtain realistic like motions [23]. These movements, in general, are captured using rt-MRI techniques [24], that reflect in the pixel values of the captured rt-MRI images. Hence, considering the pixel values in the synthesized video IF could provide realistic like articulatory movements. In the existing works on AA-video synthesis, in order to augment these movements, articulators have been considered separately. Hence, the boundaries between those articulators are required. However, we observe that automatic estimation of those boundaries in the synthesized IFs is a difficult task. We, in this work, show that those boundaries can be obtained automatically using manually annotated ATBs. Typically, the manual annotation of the ATBs takes approximately 6-10 min per IF [20]. Since the ATBs vary across the frames, the total time required to annotate is proportional to the number of IFs in a synthesized video. However, when we synthesize the articulatory videos corresponding to the test words considered in the work proposed by Desai et al. [19] with an IF sequence repository proposed by them, it is observed that a total of 1359 IFs are selected from the repository to synthesize videos containing 2123 IFs. This suggests the annotation of the ATBs for those few selected IFs would be sufficient and less time consuming.

However, it is to be noted that the IFs in the synthesized videos are not directly replaced with the IFs belonging to the selected IF sequences. Instead, the selected IFs are modified based on the methods proposed in the concatenation based approaches [19]. Hence, the ATBs belonging to the IFs in the selected IF sequences need to be synthesized according to the synthesis process, which is a challenging task. In this work, we consider the problem of automatic augmentation of the synthesized articulatory videos considering pixel values in those video IFs and the ATBs belonging to the selected IF sequences. In addition, the proposed method with the conacatenative based synthsis method, an AA-video can be obtained automatically for an expert's audio for which articulatory measurements are not available.

4. Proposed approach

Block diagram in Figure 2 shows the three major stages in the proposed method. The first stage (boundary preparation) has two steps. In the first step, we estimate the two ATBs $\mathscr{C}_1(t)$ and $\mathscr{C}_2(t) : 1 \leq t \leq T$, where T is the number of IFs in the synthesized video. In the second step, we split the two tissue regions $\mathscr{R}_1(t)$ and $\mathscr{R}_2(t)$ enclosed within $\mathscr{C}_1(t)$ and $\mathscr{C}_2(t)$ into four sub regions containing: 1) the lower-lip and the jaw $\mathscr{R}_{\text{LL-J}}(t)$, 2) the epiglottis and the tongue $\mathscr{R}_{\text{EG-T}}(t)$, 3) the upper-lip $\mathscr{R}_{\text{UL}}(t)$ 4) the palete and the velum $\mathscr{R}_{\text{P-V}}(t)$ and indicate the respective boundaries as $\mathscr{C}_r(t), r \in \{\text{LL-J}, \text{EG-T}, \text{UL}, \text{P-V}\}$ as shown in Figure 1c. In the second stage (boundary smoothing), we represent $\mathscr{C}_r(t); \forall r, t$ uniformly with fixed N points and obtain smoothed boundaries $\hat{\mathscr{C}}_r(t)$ in two steps. In the first step, we apply a low-pass filtering across the frames separately on x and y coordinate values of each location in the $\hat{\mathscr{C}}_r(t)$ to obtain smooth

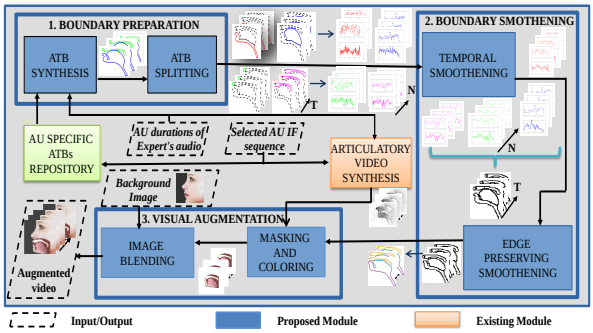

Figure 2: *Block diagram illustrating the stages involved in the proposed approach for video synthesis.*

temporal transitions in the resultant contour $\widetilde{\mathscr{C}}_r(t)$. In the second step, for every region r, we construct a gray-scale image deduced from $\widetilde{\mathscr{C}}_r(t)$ and apply an edge-preserving smoothening technique on the image to remove jagged edges. The resultant edges in each image are converted as the boundaries $\hat{\mathscr{C}}_r(t)$ for every region r. In the third stage (visual augmentation), we blend a background image and IF obtained following masking and coloring operations on the synthesized IF to obtain an augmented IF. For the coloring, we propose a set of rules specific to each region enclosed by $\hat{\mathscr{C}}_r(t)$. Finally, we incorporate the audio to obtain augmented articulatory video.

4.1. Boundary preparation

ATB synthesis: In the concatenation based synthesis, for each AU, an IF sequence of T_{AU} length is selected from an IF sequence repository and is interpolated to obtain an IF sequence of $\hat{T}_{\text{AU}}$ length. Considering these $\hat{T}_{\text{AU}}$ many IFs, the IF stitching is performed on two boundary IFs (source IFs) of two consecutive AUs to obtain one IF representing the boundary IF (target IF) between those AUs. These operations are applied on a sequence of AUs belonging to an expert's audio to obtain a synthesis video of T length. In order to augment the T many IFs in the synthesized video, we first obtain the ATBs of $\hat{T}_{\text{AU}}$ length for the altered AU IF sequence by synthesizing the ATBs belonging to the selected AU IF sequence. For this, we consider T_{AU} values of x and y coordinates corresponding to each point on the ATBs of selected AU IF sequence and interpolate those points to $\hat{T}_{\text{AU}}$ values of x and y coordinates respectively to obtain the same location on the ATBs in IF sequence of length $\hat{T}_{AU}$. This is done to achieve smooth temporal variations in the ATBs [25]. Following this, in order to achieve smooth transitions in the ATBs at the AU boundaries, we propose to synthe-

size the ATBs for the target IF, such that the entire tissue regions are enclosed by the ATBs in both the source IFs.

However, it is to be noted that ATBs consist of discrete points with a varying inter-point distance [20]. Hence the annotated points do not represent the same location on the ATB across the IFs except the start and the end points of the ATBs. In addition, the number of annotated points are not the same across the IFs [20]. These together could cause error in the proposed ATB synthesis. In order to circumvent these problems, we interpolate the points on the ATBs to obtain K equidistant points using a contour interpolation [26] with linear interpolation technique [20]. Further, by choosing a large K value, we assume that the inter-point distance could be similar across the IFs. Hence, each interpolated point would represent a spatially similar location across the frames, which, in turn, could minimize the errors in the synthesized ATBs.

ATBs splitting: The tissue color of the articulators varies according to its type and also their rate of movements. Hence, we propose to segment the articulatory region into four parts – 1) $\mathscr{R}_{\text{LL-J}}(t)$, 2) $\mathscr{R}_{\text{EG-T}}(t)$, 3) $\mathscr{R}_{\text{UL}}(t)$ 4) $\mathscr{R}_{\text{P-V}}(t)$ and augment each region according to the properties of their respective enclosed articulators. For this, we divide the regions $\mathscr{R}_1(t)$ and $\mathscr{R}_2(t)$ into those four sub-regions by manually marking following five locations on the ATBs of every IF – 1) Upper lip base (UL_{ba}) 2) start of the hard palate (HP_s) 3) hard palate end (HP_e) 4) tongue base (T_{ba}) 5) epiglottis end (EG_e). The yellow colored points in Figure 3a indicate these on the two ATBs $\mathscr{C}_1(t)$ and $\mathscr{C}_2(t)$ of an exemplary IF. In order to divide $\mathscr{R}_1(t)$, we join the points $\text{T}_{ba}, \text{EG}_e$ with a contour approximately parallel to the part of the ATB belonging to Jaw anatomical region and indicate the respective boundaries as $\mathscr{C}_{\text{LL-J}}(t)$ and $\mathscr{C}_{\text{EG-T}}(t)$.

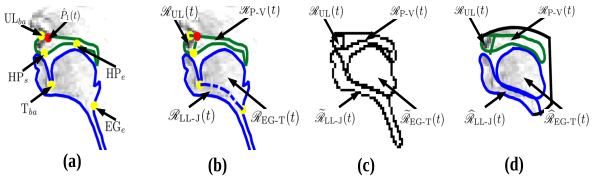

Figure 3: *Illustration of the steps in the proposed approach with an exemplary IF. a) Synthesized IF with marked and estimated points on the ATBs – $\mathscr{C}_1(t)$ (blue color), $\mathscr{C}_2(t)$ (green color). b) IF with the boundaries after ATB splitting. c) Modified IF for edge-preserving smoothening. d) Masked IF with the $\mathscr{C}_r(t)$.*

Similarly, we obtain $\mathscr{R}_{\text{UL}}(t)$ and $\mathscr{R}_{\text{P-V}}(t)$ from $\mathscr{R}_2(t)$ using the points $\text{UL}_{ba}(t), \text{HP}_s(t), \text{HP}_e(t), \text{V}_e(t)$ in two steps. In the first step, we estimate the point (red colored in 3a) $\hat{P}_1(t) = \{x^{\text{UL}_{ba}}(t) + l, y^{\text{UL}_{ba}}(t)\}$, where we found the l value empirically. In the second step, we join the points $\hat{P}_1(t)$ & $\text{HP}_s(t)$ and $\text{HP}_s(t)$ & $\text{HP}_e(t)$ with straight lines. While the annotated ATB exists in between the points $\text{HP}_s(t)$ & $\text{HP}_e(t)$, we do not consider it because we hypothesize a fixed contour between those points varies due to the errors in the annotation. We indicate the region enclosed by the contours joined by the points $\hat{P}_1(t), \text{UL}_{ba}(t), \text{HP}_s(t), \text{HP}_e(t)$ and $\hat{P}_1(t)$ as $\mathscr{R}_{\text{P-V}}(t)$ and the region enclosed by the contours joined by the points $\text{UL}_{ba}(t), \text{HP}_s(t), \hat{P}_1(t)$ and $\text{UL}_{ba}(t)$ as $\mathscr{R}_{\text{UL}}(t)$ and the respective boundaries as $\mathscr{C}_{\text{P-V}}(t)$ and $\mathscr{C}_{\text{UL}}(t)$

4.2. Boundary smoothening

In general, the manual annotation of the ATBs is performed independently across the frames without ensuring the smooth transitions across the IFs. Typically, these transitions have been observed to be low-pass in nature and depends on the type of the articulators [25]. Thus, we obtain smooth transitions in each boundary $\mathscr{C}_r(t), r \in \{\text{LL-J}, \text{EG-T}, \text{UL}, \text{P-V}\}$ across the IFs in two steps, referred as temporal smoothening. In the first step, we interpolate the discrete boundary points of the $\mathscr{C}_r(t)$ to a fixed set of N points in every IF for all r using the contour interpolation technique [26]. In the second step, we apply a low-pass filter separately on the x and y coordinates of each interpolated point across all IFs. We denote the boundary of r region after the temporal smoothening as $\widetilde{\mathscr{C}}_r(t)$.

After the temporal smoothening, we remove the rough spatial variations in the boundaries which could be due to error accumulated by temporal smoothening, ATB synthesis and annotation. For this, at each frame, we construct a gray-scale image assigning intensity value 1 to the pixel locations corresponding to the boundary points in $\widetilde{\mathscr{C}}_r(t)$ and 0 for the remaining pixel locations as shown in 3c. Following this, we apply an edge-preserving smoothening filter on the gray-scale image and consider the locations of the edges in the resultant image as the points on the smoothed boundaries $\hat{\mathscr{C}}_r(t)$. We perform this operation for each boundary of r region separately.

4.3. Visual augmentation

In the synthesized IF sequence, the regions not covered by the articulators do not convey any information, and hence, are fixed across the frames. In order to obtain better augmentation, we overlay these fixed regions with the corresponding regions taken from a realistic image belonging to a side view of a human face using an image blending technique [27]. We, in this work, consider a empirically chosen female face, back-ground image, which correctly matches the pharyngeal structures in the considered rt-MRI data. In order to perform the image blending, first, we mask the region that are not covered by the articulators. Figure 3d shows an exemplary masked IF along with the smoothed boundaries $\hat{\mathscr{C}}_r(t)$, where the masked region is obtained empirically.

Table 1: *RGB color combinations used to construct the color image for the different regions $\hat{\mathscr{R}}_r(t)$*

Region	Red (R)	Green (G)	Blue (B)
$\hat{\mathscr{R}}_{\text{LL-J}}(t)$ & $\hat{\mathscr{R}}_{\text{UL}}(t)$	235	213	208
$\hat{\mathscr{R}}_{\text{P-V}}(t)$	254	254	254
$\hat{\mathscr{R}}_{\text{EG-T}}(t)$	238	169	184
Remaining	248	128	112

Next, we color the regions ($\hat{\mathscr{R}}_r(t)$) enclosed by $\hat{\mathscr{C}}_r(t)$ using the pixel values in those regions of the masked IFs. In order to color every r region, we define the values for R-G-B color combinations in Table 1 and scale them based on the intensities computed at each pixel location (denoted by L) in that region. The intensity at the L-th pixel location is computed as $\max\left(\frac{M_r(t) - S_r^L(t)}{M_r(t)}, 1\right)$, where $M_r(t)$ and $S_r^L(t)$ are the maximum and L-th pixel location intensities in the r region of t-th synthesized IF. In the table, the R-G-B combinations for the regions $r \in \{\text{EG-T}, \text{P-V}\}$ are found empirically, however, for the regions $r \in \{\text{LL-J}, \text{UL}\}$ are found based on the R-G-B combinations belonging to the skin color in the template image. Following this, we apply mean filtering on the each r region separately. Finally, considering the colored masked IF and the back-ground image we perform image blending to obtain an augmented IF.

5. Experimental results

5.1. Experimental setup

We evaluate the quality of the AA-videos with its respective synthesized articulatory videos in a subjective manner. In the augmentation, we consider 2-d savitsky-golay filter for edge-preserving smoothening [28] and spatial mean filter of size 3 × 3. We perform the image blending following the work proposed by Perez et al. [27]. For the temporal smoothening, we consider the cut-off frequency F_c following the work proposed by Ghosh et al. [25] corresponding to α= 0.95, which denotes the total percentage of energy retrieved after low-pass filtering. In order to consider one F_c for the entire r region covered by many articulators, we use maximum value among all the F_cs belonging to those articulators. For the evaluation, we add the augmentation to the same test set of synthesized articulatory videos, which belongs to the expert's audio of 50 stimuli, considered by Desai et al. [19]. Following their work, we consider the AUs as phonemes and obtain their durations as well as corresponding selected IF sequence for each stimuli.

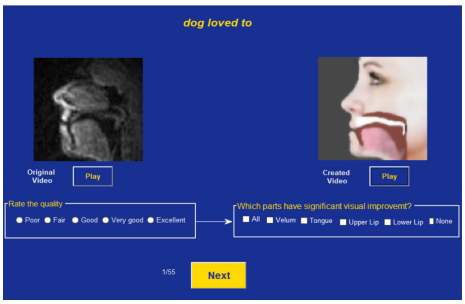

Figure 4: *Graphical user interface (GUI) used in the subjective evaluation.*

5.2. Subjective evaluation

We conduct the subjective evaluation using a set of 21 evaluators (15 males and 6 females). The evaluators are in the age group of 20 to 32 years with an average age of 22.61 years (±3.60). The evaluators are undergraduate and graduate engineering and science students. None of the evaluators has any vision problems. All the evaluators can read, write and speak English fluently.

5.2.1. Description of the evaluation set-up

In the evaluation, we present the AA-video and its respective synthesized articulatory video for each stimuli. The average duration of the videos used for evaluation is found to be 1.29 seconds (±0.39) and all the evaluators are found to be comfortable with the duration of the words used in the evaluation. Before the evaluation, we made them familiar with the articulators that are augmented in the AA-videos. We ask each evaluator to rate the visual quality of the articulatory movements in the AA-video in five categories – 1) poor 2) Fair 3) better 4) significantly better 5) excellent; compared with those in the synthesized articulatory video, when there is no significant degradation in those movements. In the case of any significant degradation compare to synthesized video, we ask them to rate the video quality as poor. In addition, we also ask them to indicate the articulators that have significant improvement in the visual appearance in following six categories – 1) all 2) velum 3) tongue 4) upper lip 5) lower lip 6) none. For this, we allow them to choose one or more categories.

This evaluation is done using a graphical user interface (GUI) developed using MATLAB R2015a as shown in Figure 4. It allows the evaluators to play the synthesized articulatory video and the AA-video separately as many times as they want. The GUI displays the audio transcription and it provides radio buttons for obtaining the evaluator ratings as well as for indicating the articulators that have significant improvement. 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.

5.2.2. Results and discussion

From the evaluator ratings, it is found that the quality of the AA-videos is 3.75 (±1.03) when averaged across all the 20 evaluators and all 50 stimuli. This indicates that the visual quality of the articulators in the AA-videos is significantly better than those in the synthesized video[1]. The highest and least ratings are found to be 4.14 and 2.62 when averaged across the ratings belonging to each word. Those ratings belong to the words "criss-crossed" and "subdued" consists of 9 and 7 AUs respectively. Further, it is also found that the average rating of the words "dessert", "tycoons" and "eleven" is 4.10 and the word "accomplish" is 2.95, which are close to the highest and the lowest ratings respectively. These words consists of a total of 6, 6, 7 and 8 AUs respectively. Moreover, we observe that there is no significant pattern between obtained average ratings and the number of AUs. This indicates that the AA-video quality is independent of number of AU boundaries in the stimuli.

Considering the total of 1050 (50×21) combinations of videos evaluated by all the 21 evaluators across 50 words, it is interesting to observe that the evaluators rate the AA-video quality as poor only in 38 number of comparisons. This indicates that the articulatory movements in AA-videos are not significantly degraded compared to the synthesized videos in most of the cases. Similarly, comparing the evaluation on articulator's visual appearance across these 1050 comparisons, it is found that all the articulators in AA-videos are significantly more visually appealing in 555 comparisons. Moreover, in only 28 number of comparisons all the articulators are chosen to improve the visual appearance to a significant level. In the remaining cases, a subset containing the combination of the articulators are found to be visually more appealing at a significant level. Among all the articulators, the upper lip has been selected in the least number of comparisons, which is 174. This could be because, for the upper lip, the number of annotated ATB points are a few; hence, a large percentage of its boundary is estimated. Thus, this causes more boundary errors leading to poor quality in the visual appearance.

6. Conclusion

We propose a method to augment a concatenation based synthesized articulatory video of an audio, for which the articulatory data is not available. The proposed method augments the videos considering synthesized IF's ATBs and the pixel values in the selected IF regions consists of articulators. While the ATBs are not directly available for the synthesized IF, we propose to synthesize those from the ATBs belonging to an IF sequence repository used in the concatenation based synthesis. Experiments with synthesized articulatory videos deduced using MRI-TIMIT containing rt-MRI videos, following subjective evaluation, reveal that the quality of articulatory movements in the augmented videos are significantly more visually appealing than the synthesized videos. Further investigations are required to develop better techniques for ATB synthesis to generalize well without considering the ATBs repository.

[1] Videos are available at https://spire.ee.iisc.ac.in/spire/software.php

7. References

[1] C. Yarra, O. D. Deshmukh, and P. K. 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] S. Nagesh, C. Yarra, O. D. Deshmukh, and P. K. 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] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection," *Speech Communication*, vol. 78, pp. 62–71, 2016.

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

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

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

[7] P. Badin, Y. Tarabalka, F. Elisei, and G. 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] P. Badin, F. Elisei, G. Bailly, and Y. 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] P. Badin, A. B. Youssef, G. Bailly, F. Elisei, and T. 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] O. 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] T. Hueber, G. Chollet, B. Denby, and M. 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] A. 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] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. 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] D. W. Massaro and J. 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] G. Bailly, P. Badin, D. Beautemps, and F. 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] T. Hueber, "Ultraspeech-player: intuitive visualization of ultrasound articulatory data for speech therapy and pronunciation training." *Proceedings of Interspeech*, pp. 752–753, 2013.

[17] B. J. Kröger, V. Graf-Borttscheller, and A. 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] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," *IEEE Signal Processing Magazine*, vol. 25, no. 3, 2008.

[19] U. Desai, C. Yarra, and P. K. Ghosh, "Concatenative articulatory video synthesis using real-time MRI data for spoken language training," *Accepted in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. –, 2018.

[20] A. K. Pattem, A. Illa, A. Afshan, and P. K. Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," *Computer Speech & Language*, vol. 47, pp. 157–174, 2018.

[21] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time MRI articulatory corpus for speech research," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[22] F. Bellard, M. Niedermayer *et al.*, "FFmpeg," *Available from: http://ffmpeg. org, last accessed on 26-10-2017*.

[23] O. Bimber, L. M. Encarnação, and D. Schmalstieg, "The virtual showcase as a new platform for augmented reality digital storytelling," in *Proceedings of the workshop on Virtual environments 2003*. ACM, 2003, pp. 87–95.

[24] A. Schiffenbauer, "Imaging: seeing muscle in new ways," *Current opinion in rheumatology*, vol. 26, no. 6, p. 712, 2014.

[25] P. K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.

[26] J. D'Errico, "Distance based interpolation along a general curve in space," *Mathworks, Available online: https://in.mathworks.com/matlabcentral/fileexchange/34874-interparc, last accessed on 23–03–18*, 2012.

[27] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Transactions on graphics (TOG)*, vol. 22, no. 3, pp. 313–318, 2003.

[28] S. Orfanidis, *Introduction to signal processing*. Pearson Education, 2010.