2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

The effect of time on the automated detection of the pharyngeal phase in videofluoroscopic swallowing studies

Andrea Bandini, *Member, IEEE,* and Catriona M. Steele

***Abstract*— Convolutional Neural Networks (CNNs) have recently been proposed to automatically detect the pharyngeal phase in videofluoroscopic swallowing studies (VFSS). However, there is a lack of consensus regarding the best algorithmic strategy to adopt for segmenting this important yet rapid phase of the swallow. Moreover, additional information is needed to understand how small the detection error should be, in view of translating this approach for use in clinical practice. In this manuscript we compare multiple CNN-based algorithms for detecting the pharyngeal phase in VFSS bolus-level clips, specifically looking at 2DCNN and 3DCNN approaches with different temporal windows as input. Our results showed that a 2DCNN analysis on 3-frame windows outperformed both frame-by-frame approaches and 3DCNNs. We also demonstrated that the detection accuracy of the pharyngeal phase is very close to the clinical gold standard (i.e., trained clinical raters). These results demonstrate the feasibility of deep learning-based algorithms for developing intelligent approaches to automatically support clinicians in the analysis of VFSS data.**

***Clinical relevance*— Accurate and reliable segmentation of the pharyngeal phase will support clinicians by reducing the time needed for rating VFSS data. Moreover, automatic detection of this phase can be seen as a foundation for building novel and intelligent approaches to detect clinical features of interest in VFSS, such as the presence of penetration-aspiration.**

I. INTRODUCTION

The videofluoroscopic swallowing study (VFSS) is the gold standard technique for clinical assessment of dysphagia (i.e., swallowing impairment) [1], [2]. VFSS is an imaging technique that involves the recording of the head and neck structures using an X-ray machine (the fluoroscope), usually in lateral view, during the swallowing of food items mixed with a radiocontrast agent such as barium sulfate. With this technique it is possible to observe bolus motion in real-time, as well as movement of the anatomical structures responsible for swallowing. Swallowing can be divided into three phases: 1) oral phase – the food or liquid is processed in the oral cavity and delivered into the pharynx by the tongue; 2) pharyngeal phase – the bolus is transported through the pharynx and upper esophageal sphincter (UES) by a series of coordinated movements, which include closure of the entrance to the airway to prevent penetration-aspiration; and 3) esophageal phase – the bolus passes through the esophagus and travels towards the stomach [3], [4].

Accurate clinical interpretation of VFSS is of paramount importance for detecting penetration-aspiration (i.e., airway invasion), as well as for planning appropriate interventions, such as the use of thickened drinks and texture-modified foods to reduce the risk of choking and aspiration [5]. To this end, standardized protocols have been developed and validated to detect physiological parameters related to the swallow and help clinicians with the decisions [6]. One of the most recent methods is the Analysis of Swallowing Physiology: Events, Kinematics and Timing (ASPEKT) [7], which provides clinicians with a rich and modular approach for rating VFSS data by detecting temporal, kinematic, and geometrical parameters pertaining to swallowing physiology.

Although standardized rating protocols have been heavily used for the assessment of dysphagia [6], there is an urgent need for video analysis algorithms that are able, at least in part, to automate VFSS rating, as the bulk of the work still relies on the manual annotation of temporal events and anatomical landmarks [7]. In fact, the vast amount of manual labour results in analysis times that are hugely disproportionate to the relatively short duration of the video clips (usually a few seconds).

With the advent of deep learning, novel and accurate algorithms for video analysis have been proposed to solve problems in the field of swallowing science. Deep learning has been applied to automatically detect and track the hyoid bone in VFSS [8], [9], to segment the bolus contour during the swallow [10], and to identify the pharyngeal phase in the swallow recordings [4], [11], [12]. The latter application is of particular interest, as the pharyngeal phase constitutes the time interval that contains most of the clinical parameters of interest [7]. Specifically, two types of video classification approaches have been proposed so far: 3D convolution neural networks (3DCNN) for classifying sliding temporal windows of consecutive frames as belonging either to the pharyngeal or non-pharyngeal phase [4], [11], and the use of frame-by-frame approaches that classify single video frames with 2DCNNs [12]. To the best of our knowledge, no previous studies have compared these two approaches on the same dataset. Nor has it been indicated how small the detection error should be to be considered acceptable for a clinical application. This information is essential to develop and deploy novel algorithms for the automatic interpretation of VFSS data.

Hence, the objectives of this paper are: 1) to determine the effect of temporal information on the automatic detection

*This work was supported by an R01 grant (NIDCD 011020) to the second author

Andrea Bandini and Catriona M. Steele are with KITE Research Institute – Toronto Rehabilitation Institute, University Health Network, Toronto, ON M5G 2A2, Canada andrea.bandini@ieee.org

Catriona M. Steele is also with the Rehabilitation Sciences Institute, Temerty Faculty of Medicine, University of Toronto, Ontario, Canada catriona.steele@uhn.ca

of the pharyngeal phase in VFSS, specifically comparing 2DCNN- and 3DCNN-based approaches for detecting this phase in bolus-level VFSS clips (i.e., is the cost of using a 3DCNN justified by higher performance?); and 2) to compare the accuracy in detecting these events with manual inter-rater agreement, to understand how far we are from deploying such video-based approaches into real-world applications (i.e., is automated detection of the pharyngeal phase as accurate as the clinical gold-standard?).

II. MATERIALS AND METHODS

A. Data Collection

The study was approved by the Research Ethics Boards at UHN – Toronto Rehabilitation Institute. All participants signed informed consent according to the requirements of the Declaration of Helsinki. Seventy-eight healthy participants (39 male, 39 female, mean age 50.3 $\pm$ 19.0 years old) with no history of swallowing, motor speech, gastroesophageal, or neurological disorders were included in the study. Twenty-seven low-concentration barium (20% w/v) stimuli were prepared and administered to each participant. Stimuli were prepared using bottled water and powdered barium sulfate (Bracco Diagnostics E-Z-PAQUE, 96% w/w) in five different consistencies: thin, slightly thick, mildly thick, moderately thick and extremely thick [5]. With the exception of thin boluses, two types of thickeners (Nestlé Resource ThickenUp Clear and Nestlé Resource ThickenUp) were used. For each combination of consistency and thickener, three boluses were swallowed by each participant. VFSS recordings were conducted in lateral projection and stored on a KayPENTAX Digital Swallow Workstation at 30 frames per per second and 720$\times$480 pixel resolution.

B. Clinical rating and pre-processing

VFSS recordings were manually split into bolus level videos and randomly assigned to two raters, who identified the two time points that delimited the pharyngeal phase, namely the bolus pass mandible (*BPM*) frame and the upper esophageal sphincter closure (*UESC*) frame (see Figure 1). The *BPM* frame is defined as "*the first frame where the leading edge of the bolus touches or crosses the shadow of the ramus of the mandible*", whereas the *UESC* frame is "*the first frame where the UES achieves closure behind the bolus tail*". Any discrepancies were resolved through a consensus meeting with a third rater [7]. Resolved values for *BPM* and *UESC* frames were considered as the ground truth for training and testing the deep learning algorithms.

Only clips with single-swallow boluses were considered for this study. Additionally, video-clips whose resolved *BPM* and *UESC* frames were deemed unratable by the raters (i.e., due to occluded images, poor quality, etc.) were excluded. The final dataset used for the analysis was composed of 1245 video clips from 59 participants. To measure inter-rater agreement, we used the Pearson's correlation coefficient (r) between the two raters (prior to discrepancy resolution) and the percentage of video clips for which disagreement was less than or equal to three frames (*P3*). These values indicated excellent inter-rater agreement (Table I).

Each video clip was split into separate grayscale frames, obtaining a dataset of 185,025 frames. A squared region of interest that included the main anatomical regions was cropped from the center of the frame and resized to 224$\times$224 pixels. Finally, a contrast-limited adaptive histogram equalization was performed to improve contrast of the anatomical structures without amplifying the noise [14]. Sample frames extracted from our dataset are shown in Figure 1.

TABLE I
INTER-RATER AGREEMENT CALCULATED ON THE DATASET OF 1245 VIDEO CLIPS.

	r	*P3* (%)
BPM frame	0.951	89.08
UESC frame	0.996	92.69

C. Automated Video Segmentation

Similar to [4], [11], [12], the problem of detecting the pharyngeal phase was tackled as a binary classification task. To automatically detect the two events of interest (i.e., *BPM* and *UESC* frames), all frames of each video clip were assigned to one of two classes, namely the pharyngeal phase (*PP*) class and the non-pharyngeal phase (*NP*) class, which included frames from the oral and esophageal phases. Frames were assigned to the *PP* class if they were between *BPM* and *UESC* frames, otherwise they were labelled as *NP*. The dataset was randomly split into training (752 clips from 38 participants for a total of 109,203 frames), validation (201 clips from 9 participants, equal to 29,491 frames), and test (292 clips from 12 participants equal to 47,331 frames) sets.

In order to investigate the effect of time in the detection of *BPM* and *UESC* frames, we conducted three tests using three different CNNs.

1) Test 1 - 2DCNN with 1 frame as input: In this test, each grayscale frame was considered independent from the preceding and subsequent frames, as proposed by Lee *et al.* [12]. We designed a custom 2DCNN with 4 convolutional layers, 2 fully connected layers with ReLU activation, and a softmax layer with 2 output units (see Table II). The size of convolutional kernel was 3$\times$3 with ReLU activation, whereas maximum pooling was performed over 2$\times$2 windows with stride equal to 2.

2) Test 2 - 2DCNN with 3 frames as input: In this test we used the same architecture developed for Test 1, with the only difference that the input was composed of stacks of 3 consecutive frames. This temporal window was advanced 1 frame at a time, thus generating an overlap of 2 frames between two consecutive time windows.

3) Test 3 - 3DCNN with 8 frames as input: In this test, the input was composed of sequences of 8 consecutive frames that were passed to a custom 3DCNN. The temporal window was advanced 4 frames at a time (i.e., 50% overlap between two consecutive samples). The architecture is similar to the

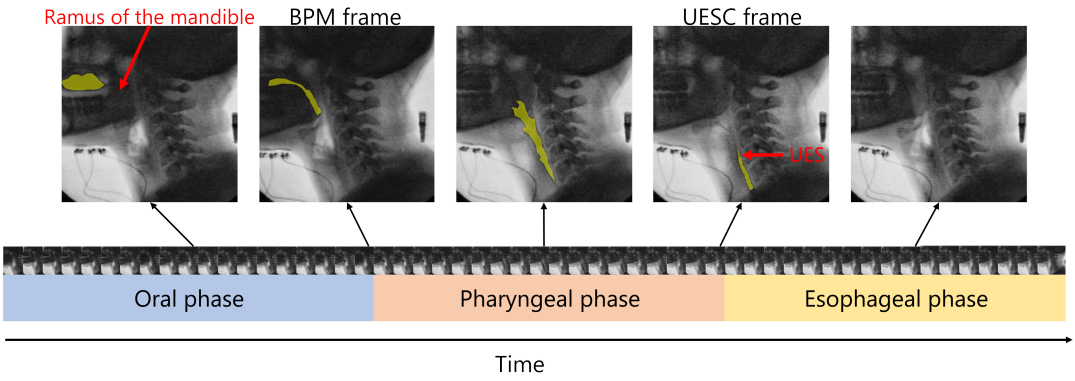

Fig. 1. Example of VFSS clip with sample frames extracted from the three phases of swallowing. The bolus is highlighted in yellow.

previous one, with the only difference that all 2D convolutional and max-pooling layers were transformed into 3D convolutions and 3D max-pooling layers (i.e. convolutional kernel size: 3×3×3; max pooling volume: 2×2×2). The fully connected and output layers remained the same as with the other two architectures (see Table II).

TABLE II

CNN ARCHITECTURES USED IN THE EXPERIMENTS. THE OUTPUT SHAPE OF EACH LAYER IS REPORTED IN PARENTHESES (CONV: CONVOLUTIONAL LAYERS; MAX-POOL: MAX POOLING LAYERS; FC: FULLY CONNECTED LAYERS).

2DCNN 1 frame	2DCNN 3 frames	3DCNN 8 frames
Input (224,224,1)	Input (224,224,3)	Input (8,224,224,1)
2D conv (224,224,4)		3D conv (8,224,224,4)
2D conv (224,224,4)		3D conv (8,224,224,4)
2D max-pool (112,112,4)		3D max-pool (4,112,112,4)
2D conv (112,112,8)		3D conv (4,112,112,8)
2D conv (112,112,8)		3D conv (4,112,112,8)
2D max-pool (56,56,8)		3D max-pool (2,56,56,8)
2D conv (56,56,16)		3D conv (2,56,56,16)
2D conv (56,56,16)		3D conv (2,56,56,16)
2D max-pool (28,28,16)		3D max-pool (1,28,28,16)
2D conv (28,28,32)		3D conv (1,28,28,32)
2D conv (28,28,32)		3D conv (1,28,28,32)
2D max-pool (14,14,32)		3D max-pool (1,14,14,32)
FC (128)		FC (128)
FC (64)		FC (64)
Softmax (2)		Softmax (2)

4) Network Training and Hyperparameters: Each architecture was trained from scratch for 100 epochs. Classification accuracy and loss obtained on the validation set were used to determine the best training hyperparameters. Specifically, for the 2DCNNs (Tests 1 and 2) the initial learning rate was set to 1e-3 and halved every 5 epochs, using a batch size of 8. The 3DCNN (Test 3) was trained using an initial learning rate of 1e-5 and halved every 10 epochs, with batch size of 4. For all models we used categorical cross-entropy loss and ADAM optimizer.

5) Performance Evaluation: For each bolus-level clip, the *BPM* frame was determined as the first frame with predicted class equal to *PP* that was followed by at least 3 consecutive *PP* frames, whereas the *UESC* frame was identified as the last *PP* frame that was preceded by at least 3 consecutive *PP* frames. This choice made the estimation robust to the presence of short and isolated windows of *PP* frames. All architectures were compared on the test set using the percentage of video-clips for which the *BPM* and *UESC* frames were predicted within 3 frames of error from the ground truth ($P3_{BPM}$ and $P3_{UESC}$). Moreover, we used as a baseline a Naïve model that used the average values of *BPM* and *UESC* frames in the training set – normalized with respect to the video duration – to predict the *BPM* and *UESC* frames on the test set.

III. RESULTS

Results are reported in Table III. All three CNNs yielded performance well above the baseline results obtained with the Naïve model. In general, the best detection results were obtained using the 2DCNN architecture with 3 consecutive frames as input. Performance in detecting the *BPM* frame was below the inter-rater agreement for all three models, whereas the 2DCNN approaches for the *UESC* frame produced values of $P3_{UESC}$ (Tests 1 and 2) higher than the inter-rater agreement obtained on the test set.

TABLE III

PHARYNGEAL PHASE DETECTION RESULTS OBTAINED ON THE TEST SET ($P3_{tot}$: PERCENTAGE OF VIDEOS FOR WHICH BOTH EVENTS – *BPM* AND *UESC* FRAMES – WERE PREDICTED WITH LESS THAN 3 FRAMES OF ERROR).

	$P3_{tot}$ (%)	$P3_{BPM}$ (%)	$P3_{UESC}$ (%)
Naïve model	14.04	13.70	14.38
Test1 (2DCNN–1 frame)	83.39	75.34	91.44
Test2 (2DCNN–3 frames)	87.16	81.16	93.15
Test3 (3DCNN–8 frames)	60.62	67.12	63.87
Inter-rater agreement obtained on the test set	88.36	86.99	89.73

IV. DISCUSSION

We demonstrated that CNN-based approaches are able to detect the *BPM* and *UESC* frames in VFSS clips with high accuracy. Our work adds further evidence to the recent results obtained by Lee *et al.* [4], [11], [12]. Specifically, our results suggest that the temporal information used as input to the CNNs has an impact on the performance of pharyngeal phase detection. The highest detection accuracy was obtained with a 2DCNN architecture with 3 consecutive frames as input. This strategy outperformed both the frame-by-frame approach (i.e., 2DCNN with 1 frame as input), and the 3DCNN-based method with 8 frames as input. Thus, the temporal extension of the analysis window might have an adverse effect on classification results. Specifically, using 8-frame analysis windows resulted in a loss of performance in capturing the right instants corresponding to the beginning and end of the pharyngeal phase (i.e., *BPM* and *UESC* frames, respectively). This result is likely due to the fact that the two events of interest happen within a short time interval, which is usually shorter than 1 second. Thus, the use of 2DCNNs, as recently proposed by Lee *et al.* [12], would be preferable for this application.

Looking at the $P3_{tot}$ values (Table III), we found that the Test 2 yielded results very close to the human inter-rater agreement (87.16% vs 88.36%). Upon closer inspection, we can see that both 2DCNN approaches (Tests 1 and 2) were able to detect the *UESC* frame with accuracy higher than the inter-rater agreement. Thus, at least for the *UESC* frame detection, an automatic video segmentation algorithm is as accurate as the trained human observer. The same cannot be said for the *BPM* frame, as the best $P3_{BPM}$ value – obtained during Test 2 – is below the inter-rater agreement (81.16% vs 86.99%). This lower performance can be explained by the fact that in some cases the *BPM* frame is identified in correspondence with premature spill of bolus into the pharynx, which might not be captured accurately by the automated approach. Thus, additional data and results are needed, in order to look more closely at this issue and improve *BPM* frame detection.

The main limitation of this study is the inclusion of only healthy participants. Future work will focus on expanding the dataset and validating the algorithm on individuals with swallowing problems, for example due to degenerative diseases or post-stroke. Moreover, an interesting development will be the comparison of our custom CNNs with popular architectures (e.g., VGG16, ResNets, Inceptions) [13], as well as the implementation of recurrent layers [14] for improving the temporal segmentation of VFSS clips.

V. CONCLUSIONS

For the first time, we compared 2DCNN and 3DCNN architectures for detecting the pharyngeal phase in VFSS, demonstrating that 2DCNNs with short temporal windows as input (i.e., 3 frames) provide better results than 3DCNNs or frame-by-frame approaches. Our results demonstrated that the automatic prediction of the pharyngeal phase can be performed with accuracy very close to the gold standard (i.e., trained clinical rater). However, further developments are needed to improve the detection performance as far as the beginning of this phase (i.e., *BPM* frame) is concerned. We are confident that with the expansion of the dataset and the implementation of other deep learning architectures (e.g., more powerful CNNs and recurrent neural networks), we will soon be able to translate this technology into clinical practice, to support clinicians with the manual rating of VFSS data.

ACKNOWLEDGMENT

The authors would like to thank all members of the Steele Swallowing Lab who participated in VFSS rating.

REFERENCES

[1] M. M. Costa, "Videofluoroscopy: the gold standard exam for studying swallowing and its dysfunction," *Arquivos de Gastroenterologia*, vol. 47, no. 4, pp. 327–328, 2010.

[2] E. Sejdic, G. A. Malandraki and J. L. Coyle, "Computational Deglutition: Using Signal- and Image-Processing Methods to Understand Swallowing and Associated Disorders [Life Sciences]," in *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 138-146, Jan. 2019

[3] K. Panara; E. Ramezanpour Ahangar; D. Padalia.; "Physiology, Swallowing," National Center for Biotechnology Information. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31082115/. [Accessed: 16-Apr-2021].

[4] J. T. Lee, E. Park, and T.-D. Jung, "Automatic Detection of the Pharyngeal Phase in Raw Videos for the Videofluoroscopic Swallowing Study Using Efficient Data Collection and 3D Convolutional Networks †," *Sensors*, vol. 19, no. 18, p. 3873, 2019.

[5] J. A. Cichero, P. Lam, C. M. Steele, B. Hanson, J. Chen, R. O. Dantas, J. Duivestein, J. Kayashita, C. Lecko, J. Murray, M. Pillay, L. Riquelme, and S. Stanschus, "Development of International Terminology and Definitions for Texture-Modified Foods and Thickened Fluids Used in Dysphagia Management: The IDDSI Framework," *Dysphagia*, vol. 32, no. 2, pp. 293–314, 2016.

[6] B. Martin-Harris, M. B. Brodsky, Y. Michel, D. O. Castell, M. Schleicher, J. Sandidge, R. Maxwell, and J. Blair, "MBS Measurement Tool for Swallow Impairment—MBSImp: Establishing a Standard," *Dysphagia*, vol. 23, no. 4, pp. 392–405, 2008.

[7] C. M. Steele, M. Peladeau-Pigeon, C. A. Barbon, B. T. Guida, A. M. Namasivayam-MacDonald, W. V. Nascimento, S. Smaoui, M. S. Tapson, T. J. Valenzano, A. A. Waito, and T. S. Wolkin, "Reference Values for Healthy Swallowing Across the Range From Thin to Extremely Thick Liquids," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 5, pp. 1338–1363, 2019.

[8] Z. Zhang, J. L. Coyle, and E. Sejdić, "Automatic hyoid bone detection in fluoroscopic images using deep learning," *Scientific Reports*, vol. 8, no. 1, 2018.

[9] D. Lee, W. H. Lee, H. G. Seo, B. -M. Oh, J. C. Lee and H. C. Kim, "Online Learning for the Hyoid Bone Tracking During Swallowing With Neck Movement Adjustment Using Semantic Segmentation," *IEEE Access*, vol. 8, pp. 157451-157461, 2020.

[10] H. Caliskan, A. S. Mahoney, J. L. Coyle and E. Sejdić, "Automated Bolus Detection in Videofluoroscopic Images of Swallowing Using Mask-RCNN," *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada, 2020, pp. 2173-2177.

[11] J. T. Lee, E. Park, J.-M. Hwang, T.-D. Jung, and D. Park, "Machine learning analysis to automatically measure response time of pharyngeal swallowing reflex in videofluoroscopic swallowing study," *Scientific Reports*, vol. 10, no. 1, 2020.

[12] K.-S. Lee, E. Lee, B. Choi, and S.-B. Pyun, "Automatic Pharyngeal Phase Recognition in Untrimmed Videofluoroscopic Swallowing Study Using Transfer Learning with Deep Convolutional Neural Networks," *Diagnostics*, vol. 11, no. 2, p. 300, 2021.

[13] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," Artificial Intelligence Review, vol. 53, no. 8, pp. 5455–5516, 2020.

[14] P. Wilhelm, J. M. Reinhardt, and D. Van Daele, "A Deep Learning Approach to Video Fluoroscopic Swallowing Exam Classification," *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020.