ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747732

SEGNET-BASED DEEP REPRESENTATION LEARNING FOR DYSPHAGIA CLASSIFICATION

Siddharth Subramani[⋆], *Achuth Rao M V*[⋆], *Anwesha Roy*[⋆], *Prasanna Suresh Hegde*[†], *Prasanta Kumar Ghosh*[⋆]

[⋆]Department of Electrical Engineering, Indian Institute of Science, Bangalore-560 012, India
[†]Dept. of Head and Neck Surgery, Health Care Global Enterprises Ltd Bangalore-560 002, India

ABSTRACT

Swallowing disorders, broadly known as Dysphagia, are difficulties in the process of swallowing food. Many currently available methods for classifying healthy and dysphagic swallows typically use hand-picked acoustic features. This article presents a SegNet-based method for classifying healthy and dysphagic swallow signals by learning mel-spectrogram features. Swallow sounds were recorded from a total of 24 subjects using a microphone based cervical auscultation (CA) system. Each subject swallowed multiple samples of water of volumes $5ml$, $10ml$ and $15ml$, and also performed multiple dry swallows. The experiments investigated the significance of temporal structures in the SegNet-learnt representations. The classification performance was evaluated at different model depths in order to identify the optimum feature time-scale that maximized the classification performance. The proposed method was found to be more robust to variations in the signatures of swallow signals across multiple volumes of water, against a baseline method across a single volume of water. The best performing model yielded a mean test F1-score of 80.13% (±4.62%) in a 5-fold cross validation setup.

Index Terms— Cervical Auscultation, Dysphagia Classification, SegNet, Two-step training

1. INTRODUCTION

The process of swallowing (deglutition) in humans involves passing the saliva mixed food bolus from the mouth to the stomach through peristalsis in the esophagus. This is an intricate process coordinated by around 30 muscles controlled by the cerebral cortex of the human brain. Since the pharynx is the common passage to air and food, the epiglottis is shut in the pharyngeal phase to stop food from entering the airway. This would otherwise result in pulmonary aspiration [1]. Any difficulty or the need for increased effort in swallowing is referred to as dysphagia. Causes of dysphagia include weakening of muscles, neurological disorders like multiple sclerosis, stroke, Parkinson's disease, and obstructions due to laryngeal, esophageal or head & neck cancer. Dysphagia can cause illness pertaining to aspiration pneumonia, malnutrition, weight loss, dehydration and choking [2]. Common clinical interventions that diagnose dysphagia include invasive methods like fiber-optic endoscopy, videofluoroscopy, functional magnetic resonance imaging and non-invasive methods like surface electromyography and cervical auscultation (CA) [3]. Of such methods, CA, which uses a simple acquisition device setup to record swallow signals, has been shown to produce representations of swallows similar to other clinical methods [4]. Identifying dysphagia is a task of foremost importance due to its detrimental consequences. Reducing manual efforts and possible human errors by assisting clinical experts in identifying dysphagia helps in improving the overall efficacy of the diagnostic process. This can be achieved through automating a system to learn features for characterizing and classifying dysphagic swallows.

Previous works in the literature, using CA, have characterized the process of swallow using either accelerometer vibrations or microphone sounds, but largely using the former. Contrast to popular usage of accelerometer vibrations in the literature, for swallow signal characterization and classification (like in works [5, 6, 7, 8]), Cichero et al. [9] concluded that swallow signals acquired by electret microphones are more informative than accelerometer vibrations. This was attributed to their improved signal-to-noise ratio characteristics and resistance against ambient background noise. The works by Dudik et al. [10] and Movahedi et al. [11] compared swallow sounds and vibrations across multiple bolus consistencies. It was found that the time-frequency and energy characteristics of the signals acquired using each device differed for the healthy and dysphagic classes. For the same reason, the representations by the accelometer and the microphone were deemed to be non-interchangeable. Miyagi et al. [12] used a Radial Basis Function (RBF) kernel SVM classifier to classify healthy and dysphagic swallows. Their dataset comprised sounds from healthy and dysphagic subjects swallowing multiple samples of 3ml of water. Their best performing feature set pertained to peak amplitude and quartile-ratio features from different spectral regions of swallow signal spectrograms.

In works such as mentioned above, comparatively, only few have characterized dysphagic swallows using microphone recordings. In general, many works required manual engineering of signal features to an extent. Also, not many works in the literature have been able to conclusively define spectro-temporal variations of swallow sounds, especially in dysphagic sounds. This makes manual selection of acoustic features difficult. Thus, automatic feature learning and dysphagia identification will help in building acoustic representations of swallow signals and in reducing manual efforts and errors. This two-step process is achieved in the current work using an autoencoder framework for learning features from mel-spectrograms and then training a classifier using the learnt features. In this work no procedures for swallow phase segmentation were employed and features were learnt from each entire swallow recording (ie., without explicitly learning potentially influential features from individual swallow phases). A microphone based cervical auscultation setup was used to acquire swallow signals from healthy and dysphagic subjects. Since the number of swallow signals in the control and patient classes used in this work was unbalanced, F1-score was used as the evaluation metric. The subjects were split into 5 distinct folds comprising train, validation and test sets. The proposed method was able to achieve an F1-score of 80.13% (±4.62%) across the 5 folds (over multiple bolus volumes).

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

2. DATASET

The dataset for this work was collected at HealthCare Global Enterprises Limited, Bangalore, India. Consent was obtained from all subjects whose swallow signals were recorded. Swallow sounds from 14 healthy controls (21 - 36 years) and 10 patients (34 - 75 years) were obtained using a microphone-based CA setup. It has been discussed in [13] that the mass and strength of muscles coordinating the process of swallow deteriorate with age. This induces an unintentional fatigue in chewing and swallowing foods and presents an increased risk of choking, especially among people older than 65 years of age. Hence, in this study, healthy swallows were recorded from individuals who were considerably younger than those in the dysphagic group.

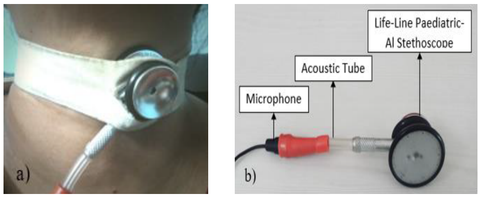

Fig. 1: (a) Cervical auscultation device attached to the neck (b) Components of the cervical auscultation device

The swallow signals were recorded in a well-illuminated and a relatively quiet environment. The data collection setup used in this work (shown in Fig. 1) was the same as that used by [14]. The Life-Line Paediatric-Al Stethoscope was used to record swallow sounds. The output from this device was connected via an acoustic tube to Sorella'z portable 3.5mm microphone (frequency range of 30Hz-15kHz, sensitivity value of -52dB (5dB tolerance) and an impedance value of 2.2kΩ). After each subject was seated, the auscultation device was secured to their neck region corresponding to the lateral border of the trachea, ensuring no obstruction to normal breathing and swallowing in the subjects. For CA, this region which is inferior to the cricoid cartilage encircling the trachea was concluded as the optimal site for recording swallow sounds by Pan Q. et al. [15].

Each subject in their resting state swallowed separate samples of 5ml, 10ml and 15ml of water. The subjects also performed dry swallows without any water. In the dry swallow the subjects might have swallowed saliva. Boluses of volume 5ml, 10ml and 15ml (measured by a syringe) were provided to the subjects in paper cups. For each swallow attempt, the subjects were prompted to swallow the water from the cup at one go. A separate mobile application was used to timestamp the onset and ending of each attempt. Swallows accompanied by nasal leakage, coughs, drooling or uttering undesirable and unexpected sounds (particularly in the control group) were discarded from the dataset. This process was repeated for 3-4 times for each volume of water with ample rest between each attempt. A rest of about 2 minutes was allocated between recording sessions for different bolus volumes. In such a manner, a total of 290 swallow signals (172 healthy and 118 dysphagic) were obtained. The average number of swallows per subject in the control group was 12.28 and that in the patients group was 11.8. PRAAT software [16] was used to process and digitally store all recorded swallows at a sampling rate of 16kHz. This in-house dataset will henceforth be referred to as InD in this article.

3. PROPOSED METHODOLOGY

The proposed methodology involves, (a) computation of acoustic features, (b) learning acoustic representations for swallow signals and (c) classification of swallows into healthy and dysphagic groups using a binary classifier. The details of the 3 steps are discussed below.

3.1. Input features and data pre-processing

Mel-spectrogram, henceforth referred to as MSpec in this article, was computed from each swallow signal using a hamming window of length 20ms and a hop length of 2ms. The librosa python package for music and audio analysis [17] was used to compute all MSpecs. The MSpecs had different number of timesteps due to variations in the duration of swallow signals. All MSpecs were hence augmented to equal number of timesteps by padding them with zeros. Binary one-dimensional (1D) masks were created with zeros at indices corresponding to the padded portions of the MSpecs and ones at indices corresponding to the actual portions of the MSpecs. Each MSpec was multiplied with its corresponding binary mask before being fed to the neural network layers. This process ensured that the model learnt from only the actual regions of the input feature maps and not from regions that do not correspond to the actual MSpec.

3.2. Model architecture and classification techniques

SegNet [18], a deep convolutional autoencoder architecture, with 13 1D convolutional layers, was employed for the task of learning acoustic features from swallow signals. Fig. 2 outlines the model architecture and the proposed methodology. The SegNet encoder-decoder model preserves spatial information of the input MSpec by optimizing the network parameters with the aim of learning reduced-dimension bottleneck features at the encoder layers which are then used for reconstructing the same features in the layers of the decoder section. Using the features learnt by the SegNet encoder as input, three approaches for dysphagia classification were experimented -

- **Approach 1** (CNN1) - *1D-CNN Classifier:* A 1D-CNN based deep neural network was first used to perform binary classification on the input data. Here the chosen model architecture was the SegNet encoder followed by a sigmoid activation layer for classification.
- **Approach 2** (Jnt) - *Joint Training:* Using flattened representations of the encoder output, a dense layer followed by sigmoid activation was trained along with the autoencoder. This model learns by jointly optimizing the sum of the reconstruction loss at the decoder and the classification loss at the sigmoid layer.
- **Approach 3** (TsT) - *Two-step Training:* The SegNet-learnt bottleneck features were used to train a linear classifier. A Support Vector Machine (SVM) model [19] was chosen as the binary classifier due to their robustness against overfitting (when working with small datasets) and outliers in the input data.

4. RESULTS & DISCUSSIONS

4.1. Experimental Setup

MSpec features and binary masks were computed for all swallow signals. No signal pre-processing techniques were employed prior to computing MSpec from raw signals. All 24 subjects were split into train, validation and test sets. A 5-fold cross validation setup was used wherein in each fold the subjects were randomly picked

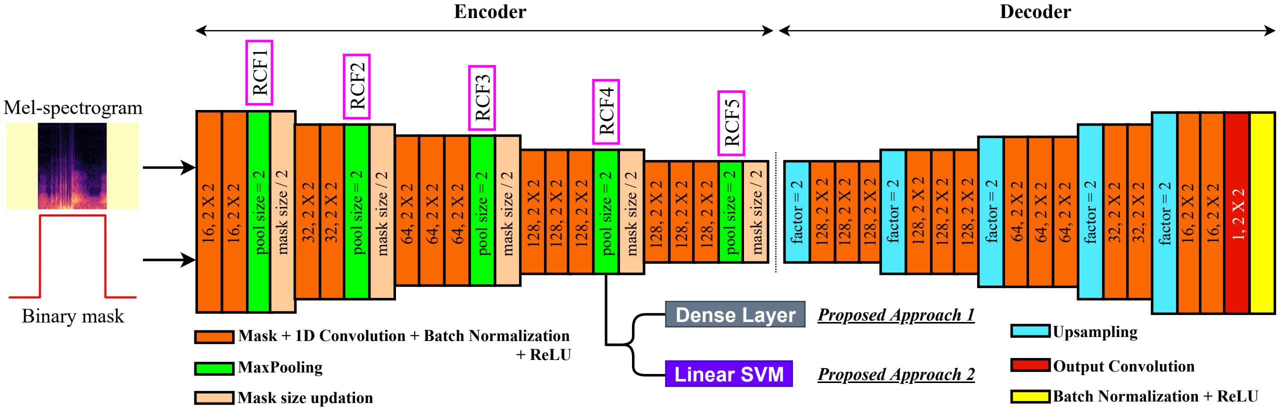

Fig. 2: An illustration of SegNet Architecture; Maxpooling layers RCFi ($i = 1, 2, 3, 4, 5$) have receptive field sizes of 3, 10, 26, 58 and 122

such that there are 2 patients for every 3 controls. In each fold, none of the train, validation and test sets had common subjects. The Adam optimizer [20] was used to train the SegNet by optimizing the mean squared error loss between the input and the reconstructed output under an early stopping criterion (with a patience of 8 training epochs) based on the mean absolute error on the validation set. The initial learning rate was set to 10^{-3}, with a decay rate of 10^{-6}. With respect to the classifier, unlike the baseline scheme, a linear kernel was used in this work since kernels such as RBF have been found to be unsuitable when the dimension of features is large [21]. While training the linear SVM in TsT, a grid search was performed (from 10^{-5} to 10^{5}, in multiples of 10) for determining the regularization parameter (C). This selection was based on the performance of the SVM classifier on the validation data. Since the healthy and dysphagic swallow classes were imbalanced, the F1-score metric [22] was used to evaluate the performance of the proposed approaches. In this article, standard deviation values are provided in brackets whenever mean F1-scores are mentioned. Sensitivity rate and specificity rate were also calculated to gain insight into the classifiers' ability in producing true predictions on the given data. The results obtained using the baseline scheme and proposed approaches are summarized in Fig. 3 and Table 1.

4.1.1. *Effect of Receptive Field Size*

It is important for the model to learn features corresponding to different phases of bolus transfer during swallowing (oral, pharyngeal and esophageal [23]), and also from regions of transition from one phase to another. This learning can be optimized through finding the right feature time-scale (i.e., the receptive field size at different depths of the SegNet encoder) that results in an encoded representation most suitable for reconstruction. Fig. 2 shows the receptive field size at different depths of the SegNet encoder architecture. In approaches CNN1, Jnt and TsT, five RCFi-trials (an RCFi-trial denotes training the binary classifier using the embedded representations obtained from the corresponding maxpooling layer denoted as RCFi in Fig. 2) were performed to find out the optimum feature time-scale.

4.2. Results with baseline scheme

The method proposed by Miyagi et al. [12] was considered as the baseline scheme here since their work also uses microphone recordings of swallow signals for dysphagia classification. The RBF-SVM model was trained (on InD) using their best performing features computed from discrete time fourier transform and spectrogram of each signal. While evaluating the performance of the baseline SVM model on InD, the validation set was clubbed with the training set. In the proposed approaches however, the training and validation sets were used to train the SegNet. The test set was kept exclusively to predicting on unseen data. This model achieved a mean test F1-score of 71.95% (±9.84%) on InD. The baseline scheme achieved a corresponding F1-score of 78.9% on their dataset that consisted of 104 randomly selected signals (at 8000Hz, from 27 healthy controls and 143 patients) in both classes. The poorer performance on InD shows that the baseline scheme suffers on larger and imbalanced classes. Though the baseline scheme produced a good sensitivity rate of 76.67% on the test dataset, it suffered from a poor specificity rate of 42.92% (ie., high false positive results). Also, since foldwise F1-scores were unavailable from the baseline work, it was not possible to further compare their dataset and InD.

4.3. Results with 1D-CNN Classifier

The 1D-CNN classifier was trained separately as a binary classifier (with no aim of input reconstruction, unlike that in approaches Jnt or TsT). In initial trials, 1D-CNN models of different architectures for the classification task were experimented with. Since all their performances were similarly poor (due to overfitting), finally a model with the architecture as mentioned in approach CNN1 (in subsection 3.2) was selected. This was done to verify if the bottleneck feature learning step (as in approaches Jnt and TsT) indeed improves the discriminative ability of the model. This model was trained under each RCFi-trial. Of all RCF-trials, RCF3 performed the best with a mean test F1 score of 68.66% (±11.94%). However, this approach, on all RCF-trials performed poorly on the classification task - in general, it showed poor sensitivity rates (ie., high false negative predictions) than the baseline scheme, as shown in Table 1. Hence, after this, approach JnT was experimented with to possibly improve model performance using joint training.

4.4. Results with Joint Training

Using approach Jnt in RCF5-trial, the mean test F1-score of 74.22% (±15.25%), across all 5 folds, was greater than the baseline value by 2.27% (absolute). These results show the capability of the proposed model to perform better on imbalanced classes. Further, the other RCF-trials were carried out in this approach. Fig. 3 summarizes the mean test F1-scores across all folds for each RCF-trial. The RCF4-trial performed the best with an average F1-score of 74.26% (±8.77%) and mean test sensitivity and specificity of 71.53% and

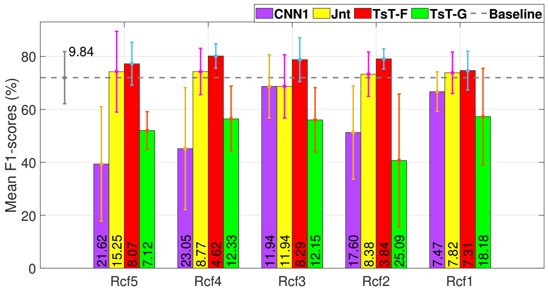

Fig. 3: Performance of baseline scheme, CNN1, Jnt and TsT (apart from error bars, corresponding standard deviation values (rounded to 2 decimal places) of mean F1-scores are displayed within each bar

67.21% respectively. Every RCF-trial however showed large variations in their fold-wise performance (high standard deviation values as illustrated in Fig. 3). This was contributed by the fact that the neural network suffered from overfitting - large variations on the performance on the validation set with every progressing epoch - which directly impacted the overall performance of the model. This presented a necessity to curb overfitting and reduce performance variation across folds. To address this problem, in the next approach (TsT), classification was performed using an SVM classifier trained on the encoder-learnt representations.

4.5. Results with Two-step Training

Here the SegNet encoder-representations were first flattened and then standardized to zero mean and unit variance before training the SVM classifier. This method of input data representation and SVM training is henceforth referred to as TsT-F in this article. The mean test F1-score of 77.21% (±8.07%) in the RCF5-trial was 2.7% greater than the baseline value. Of all methods, RCF4-trial in TsT-F performed the best with a mean test F1-score of 80.13% (±4.62%). Using the model architecture in RCF4-trial setup, the optimum receptive field size (feature time-scale value) was calculated to be 58. The mean test sensitivity and specificity values obtained from RCF4-trial was the best combination when compared to those from other methods where atleast one of the two metrics was poorer. From Fig. 3 it can be observed that the standard deviation of the fold-wise test F1-scores in RCF4-trial in TsT-F was reduced to ±4.62% from ±8.77% in the RCF4-trial of Jnt and ±9.84% in the baseline scheme. Adding to this, the average values of sensitivity and specificity rates across all RCF*i*-trials was found to be the largest when compared against all other methods listed in this work. This indicates the robustness of the features learnt using TsT-F, in general and when limited to its RCF4-trial, for classifying swallow signal characteristics from both groups. The improved results from TsT-F also show that the learning of compressed input feature space and the usage of SVM classifier help in alleviating the problem of overfitting.

Further, to comprehend the influence of temporal characteristics of the features in dysphagia classification the SVM model was also separately trained by using a global-maxpooled version of the SegNet encoder-learnt bottleneck features (approach TsT-G). In Global-Maxpooling, only the maximum value in each convolution channel is retained. In TsT-G, the model resulted in lower values of sensitivity rate, specificity rate and mean F1-scores and, higher standard deviations compared to TsT-F. This indicates that the complete absence of temporal dependencies in the bottleneck features fed to the SVM degrades the performance (especially since duration of swallow can vary) and that the flattened bottleneck features that retain temporal dependencies allow the classifier to learn more crucial characteristics for classification. Apart from having the highest mean test F1-scores, the average standard deviation of mean test F1-scores across all RCF*i*-trials in TsT-F (±6.43%) was the lowest value compared to the average standard deviation values when computed using the baseline scheme, CNN1, Jnt and TsT-G approaches. This stands in support to show that the features learnt through the Two-step Training approach achieves improved generalization in representing swallow signals and hence emerges winner in classifying healthy and dysphagic swallows.

Table 1: Mean values of sensitivity and specificity of all approaches across 5 folds

	Mean Sensitivity across folds (%)				Mean Specificity across folds (%)			
	CNN1	Jnt	TsT-F	TsT-G	CNN1	Jnt	TsT-F	TsT-G
Rcf5	31.74	71.45	79.27	48.67	74.22	67.67	61.49	45.87
Rcf4	40.09	71.53	78.24	51.85	71.75	67.21	74.38	55.43
Rcf3	24.39	68.05	78.48	50.93	73.19	60.04	69.64	57.02
Rcf2	54.29	78.96	81.62	37.31	32.92	49.02	62.81	40.61
Rcf1	64.76	74.54	73.83	57.62	58.62	65.65	66.57	39.73
Mean	43.05	72.91	**78.29**	49.28	62.14	62.92	**66.98**	47.73
	Baseline Sensitivity = 76.67				Baseline Specificity = 42.92			

5. CONCLUSIONS

This work presents a deep representation learning approach for classifying healthy and dysphagic swallows using automatically learnt representations from mel-spectrogram features. From the experiments conducted, the Two-step Training model performed the best with a mean test F1-score of 80.13%, mean test sensitivity of 78.24% and mean test specificity of 74.38%. This, in addition to helping in identifying the optimum feature time-scale for maximizing model performance, highlights the ability of the model to generalize across swallows of varying bolus volumes from the control and patient groups. The improved metrics (against the baseline scheme) indicate the robustness of the model wherein it produces lesser false predictions and hence implies better feature learning. Future work includes expanding the swallow signals dataset and adapting the proposed method to learn to categorize the severity and sub-classes of dysphagia.

6. ACKNOWLEDGMENT

The authors thank all the subjects for participating in the swallow sound data collection. The authors also thank the Pratiksha Trust and DST, Govt. of India for their support.

7. REFERENCES

[1] Takashi Nishino, “The swallowing reflex and its significance as an airway defensive reflex,” *Frontiers in physiology*, vol. 3, pp. 489, 2013.

[2] C Gallegos, E Brito-De La Fuente, P Clavé, A Costa, and G Assegehegn, “Nutritional aspects of dysphagia management,” *Advances in food and nutrition research*, vol. 81, pp. 271–318, 2017.

[3] Renée Speyer, "Oropharyngeal dysphagia: screening and assessment," *Otolaryngologic Clinics of North America*, vol. 46, no. 6, pp. 989–1008, 2013.

[4] Geovana de Paula Bolzan, Mara Keli Christmann, Luana Cristina Berwig, Cintia Conceição Costa, and Renata Mancopes Rocha, "Contribution of the cervical auscultation in clinical assessment of the oropharyngeal dysphagia," *Revista CEFAC*, vol. 15, no. 2, pp. 455–465, 2013.

[5] Lisa J Lazareck and Zahra MK Moussavi, "Classification of normal and dysphagic swallows by acoustical means," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 12, pp. 2103–2112, 2004.

[6] Azadeh Yadollahi and Zahra Moussavi, "Feature selection for swallowing sounds classification," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 3172–3175.

[7] Joon Lee, Stefanie Blain, Mike Casas, Dave Kenny, Glenn Berall, and Tom Chau, "A radial basis classifier for the automatic detection of aspiration in children with dysphagia," *Journal of NeuroEngineering and Rehabilitation*, vol. 3, no. 1, pp. 1–17, 2006.

[8] Joshua M Dudik, James L Coyle, Amro El-Jaroudi, Zhi-Hong Mao, Mingui Sun, and Ervin Sejdić, "Deep learning for classification of normal swallows in adults," *Neurocomputing*, vol. 285, pp. 1–9, 2018.

[9] Julie AY Cichero and Bruce E Murdoch, "Detection of swallowing sounds: methodology revisited," *Dysphagia*, vol. 17, no. 1, pp. 40–49, 2002.

[10] Joshua M Dudik, Atsuko Kurosu, James L Coyle, and Ervin Sejdić, "Dysphagia and its effects on swallowing sounds and vibrations in adults," *Biomedical engineering online*, vol. 17, no. 1, pp. 1–18, 2018.

[11] Faezeh Movahedi, Atsuko Kurosu, James L Coyle, Subashan Perera, and Ervin Sejdić, "A comparison between swallowing sounds and vibrations in patients with dysphagia," *Computer methods and programs in biomedicine*, vol. 144, pp. 179–187, 2017.

[12] Shigeyuki Miyagi, Syo Sugiyama, Keiko Kozawa, Sueyoshi Moritani, Shin-ichi Sakamoto, and Osamu Sakai, "Classifying dysphagic swallowing sounds with support vector machines," in *Healthcare*. Multidisciplinary Digital Publishing Institute, 2020, vol. 8, p. 103.

[13] Julie AY Cichero, "Age-related changes to eating and swallowing impact frailty: Aspiration, choking risk, modified food texture and autonomy of choice," *Geriatrics*, vol. 3, no. 4, pp. 69, 2018.

[14] Siddharth Subramani, MV Achuth Rao, Divya Giridhar, Prasanna Suresh Hegde, and Prasanta Kumar Ghosh, "Automatic classification of volumes of water using swallow sounds from cervical auscultation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1185–1189.

[15] Q Pan, Naoto Maeda, Yousuke Manda, Naoki Kodama, and Shougo Minagi, "Validation of the optimal site in the neck region for detecting swallowing sounds," *Journal of oral rehabilitation*, vol. 43, no. 11, pp. 840–846, 2016.

[16] Paul Boersma and David Weenink, "Praat: doing phonetics by computer (version 5.1. 05)[computer program]. retrieved may 1, 2009," 2009.

[17] Brian McFee, Vincent Lostanlen, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana, Kyungyun Lee, Oriol Nieto, Jack Mason, Dan Ellis, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, Keunwoo Choi, viktorandreevichmorozov, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, nullmightybofo, Darío Hereñú, Fabian-Robert Stöter, Pius Friesch, Adam Weiss, Matt Vollrath, and Taewoon Kim, "librosa/librosa: 0.8.0," July 2020.

[18] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[19] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al., "A practical guide to support vector classification," 2003.

[22] Yutaka Sasaki et al., "The truth of the F-measure," *Teach Tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.

[23] Sylvain Morinière, Patrice Beutter, and Michèle Boiron, "Sound component duration of healthy human pharyngoesophageal swallowing: a gender comparison study," *Dysphagia*, vol. 21, no. 3, pp. 175–182, 2006.