AUTOMATIC CLASSIFICATION OF VOLUMES OF WATER USING SWALLOW SOUNDS FROM CERVICAL AUSCULTATION

Siddharth Subramani†, Achuth Rao M V†, Divya Giridhar†, 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

The signatures of swallowing vary depending on the volume of bolus swallowed. Among existing instrumental methods, cervical auscultation (CA) captures the acoustic signatures of the swallow sound. Although many features present in the literature can characterize volumes of swallow using CA, they require manual annotations of the different components in the sound. In this work, a rich set of acoustic features, the ComParE 2016 acoustic feature set (OS) is used to investigate whether several temporal, spectral, vocal and source features and their functionals provide cues for volume classification. Experiments are performed with CA data from 56 subjects, with dry swallow and swallows of 2ml, 5ml, and 10ml of water. Three types of classification namely, dry-vs-2ml, dry-vs-5ml and dry-vs-10ml are performed separately to analyze characteristic features. Experiments reveal that OS, which does not require annotations, performs better than the baseline features that require annotation. Within OS, the features unrelated to voice source yield a better performance than the features related to voice source. In this subset of features, MFCC, RASTA filtered audio spectrum and RMS energy are found to be consistently the top performing features across all three types of classifications.

Index Terms— swallow sound signal, cervical auscultation, acoustic analysis, feature selection

1. INTRODUCTION

In humans, swallowing is the act of movement of food from the mouth through the pharynx to the oesophagus. The voluntarily chewed food then blends with saliva to form the food bolus. The action of swallowing requires the coordination of around 30 muscles controlled by the cortical areas of the human brain. The process of swallowing is in sync with respiration since the pharynx is the common opening to both food and oxygen. This makes the swallow a very intricate and complex process. During swallowing food, the airway is protected by the muscles of the respiratory tract, which otherwise might lead to pulmonary aspiration [1]. Swallowing disorders, generally termed as dysphagia, can be caused due to various reasons. It could be due to neurological disorders like Parkinson’s disease, stroke, head and neck cancer, amyotrophic lateral sclerosis or irregularities in the pharyngeal and oesophageal muscles. Swallow is the second step in the intake of food for the nourishment of the human body. Swallowing disorders can thus lead to ailments or conditions such as fatigue, loss of weight, pulmonary aspiration, malnutrition and choking [2]. Diagnosis of such disorders usually involves a screening test by speech language pathologists. This helps to determine the condition of dysphagia, which serves as a ground for further evaluations. Common clinical swallow assessments include videofluoroscopy, fiber-optic endoscopy, surface electromyography, cervical auscultation and ultrasonography [3, 4, 5]. Of the above mentioned methods, cervical auscultation is a non-invasive method of listening swallow sounds through a stethoscope or a microphone [3, 4]. This method agrees to a certain extent with other instrumental methods[3].

Since cervical auscultation picks up the swallow sound, the action of swallowing can be represented and characterized in terms of the swallow signal. For this, it is important to understand swallow signals and their components. Previous research works discuss about the different components of swallow signals. In [6] the optimal sites for detection of swallow sounds were discussed. The posterior inferior to the cricoid cartilage encircling the trachea was shown as the optimal region. Also, from [7] it can be seen that the electret microphone was rated as a better option for acoustic detector compared to accelerometer due to its improved resistance to ambient noise levels and signal-to-noise ratio. In [8] it was noted that the average swallow sound frequency range was 2200 Hz. The stable features of swallow signals were however found to be intensity, location and frequency of peak portion of the signal. Lazareck and Moussavi [9] performed classification of normal and dysphagic swallows using a discriminant algorithm. This was in an attempt to reduce the need for conducting videofluoroscopic studies for swallow analysis. Kamiyanagi et al. [9] measured the duration and peak intensity of the swallow signals from patients who had undergone maxillectomy. Here, the effect of expiration during swallowing was compared with the corresponding videofluoroscopic results.

Honda et al. [10] characterized the swallow signals into three components (first, second and third Swallowing Sound Wave (SSW)). Their analysis showed that a significant correlation was observed between the bolus volume and duration of second SSW and the peak intensity ratio. With their synchronized experimentation using both videofluoroscopy and acoustic analysis, they concluded that swallow signal is associated with oral, pharyngeal and re-positioning phases, each corresponding to an SSW, respectively. Giridhar et al. [11] concluded that, from among the three swallow sound waves, the peak intensity of the second swallow segment is the most significantly varying parameter across different volumes of water. Morini et al. [12] studied the duration of swallow sound components across healthy male and female individuals. They defined three main swallow components (SCs) and the intervals (I) between two successive SCs. Their results showed no statistically significant gender difference in swallow sound and in their corresponding SC’s duration. Jayatilake et al. [13] proposed a smartphone based real-time swallowing ability assessment called swallowscope for automatically recognizing dry and water swallow.

In the works from the literature mentioned above, acoustic analysis was performed using few subjects. All the features discussed in the above mentioned works required the expert annotations of the boundaries of SCs or SSWs, which could be time consuming and
From Fig. 1, a variability in the length of SC1, SC2 and SC3 across the different volumes of swallow, can be observed. This suggests that it could be difficult to characterize and develop a volume independent model. Hence, it is necessary to distinguish and learn about the features that are representative of different volumes of swallow. In the current work, the effectiveness of various acoustic features from ComParE 2016 acoustic feature set was evaluated for classifying swallowing volumes. The entire set of features has 26 feature groups. All of the features from this set were extracted from the whole signal and not from any certain components of the signals. This was termed as the OS feature set. To automatically classify healthy and dysphagic patients, it is necessary to understand the characteristics of the swallow signal and their tendency based on the volume of the bolus. Such a volume specific model can help further in classifying subjects with swallowing disorders and study their severity.

In the current study, cervical auscultation was used to study the swallow signal features for classifying dry swallow against 2ml, 5ml and 10ml water swallows, using a linear classifier. It was found that the OS features performed better than the swallow parameters which characterize different SCs as shown in Fig. 1. In particular, the OS features resulted in accuracies which were 32.88%, 15.67%, 8.35% (absolute) better than those using swallow parameters (also referred to as baseline features) for dry-vs-2ml, dry-vs-5ml and dry-vs-10ml respectively. The OS features outperformed the baseline features for all three classification tasks, particularly by a large margin for dry-vs-2ml classification. Based on feature group selection, MFCC, RASTA filtered audio spectrum and RMS energy were found to be consistently good features across three different classifications. A group of 12 features, comprising of loudness and its various robust statistics emerged from a feature selection process, were found to perform better than the 12 baseline features and not significantly different from the OS feature set.

2. DATASET

For the purpose of volume classification using swallowing sounds, 56 subjects were made to swallow water of three volumes namely, 2ml, 5ml and 10ml. Apart from these, the subjects were also made to perform dry swallow. For dry swallowing the subject is made to perform swallowing in his or her resting state, without any bolus. During this dry swallow, saliva might have been swallowed. All the data were collected from people within 20-30 years of age. Fig. 2 shows the device setup of the cervical auscultation system. The digital cervical auscultation system was installed in a luminous and a relatively noise free surrounding. The auscultation device consisted of

Fig. 1: Sample swallow signals of different volumes; regions SC1, SC2 and SC3 are annotated by the red, black and magenta lines, respectively.
feature extraction is done in two steps. In the first step, the acoustic low-level descriptors (such as MFCC, PLP etc.) are extracted using a window of length of 25ms and then the low level descriptors at each time are pooled using a variety of statistical and modulation functionals. This results in total of 6373 features for each recording. The details of the low level descriptors are given in Table 2. Broadly, the feature set can be divided into two groups (A and B) based on the features. Group A features arise largely from the vocal tract parameters whereas group B represents features from glottal excitation. These features were computed using the OpenSMILE feature extraction and audio analysis tool [15]. In this work, the features are referred to as OS.

<table>
<thead>
<tr>
<th>Clases</th>
<th>Feature name / Abbreviation / Dimension</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A</td>
<td>Loudness (LdBx3), RASTA (26), MFCC (1), RMS energy (RMSx1), Modulated Loudness (MLabsx1), ZCR (1), Band energy (BEx2), Spectral: ROP (SR)x4, Flux (SFx1), Centroid (SC)x1, Slope (SS)x4, Entropy (SE)x1, Variance (SV)x3, Harmonicity (SH)x1, Sharpness (Sh)x1</td>
<td>59 X 54 functionals + 59 delta X 46 functionals = 5900 features</td>
</tr>
<tr>
<td>Group B</td>
<td>6 features + 6 delta features</td>
<td>6 X 39 functionals + 6 delta X 39 functionals = 468 features</td>
</tr>
</tbody>
</table>

### 3.2. Baseline features

In the literature, various acoustic features of swallowing sounds were shown to significantly vary for different volumes [10, 12, 16]. Combining all the features from the literature results in a total of 12 features [10, 12, 16]: duration of swallow waves (SSW1, SSW2), duration of swallow sound components (SC1, SC2, SC3), duration of intervals between swallow components (I1, I2), duration to peak intensity (DPI), peak intensities (PI, SSW1, SSW2, PI) and the total duration (TD) of the swallow signal. These features are referred as baseline features. The calculation of these features required the manual annotation of each swallow signal into its respective components.

### 3.3. Study design 1: All acoustic features set

This study design consisted of assessing the overall performance of volume classification using the baseline features, which require annotation, versus the OS feature set which does not require annotation. The volume discrimination capacity of each feature was also studied to figure out the key features for classification.

### 3.4. Study design 2: Category-wise analysis

This study design was to assess the importance of the different categories of features in the OS features. The first study was to find out if the vocal tract related (Group A) or glottal excitation related (Group B) features are important for classifying the volume. Following this, the importance of each low level feature category was examined and the performance of each low level feature with all its functionals, of the OS feature set, was assessed.

### 3.5. Study design 3: Feature selection

Even though each category of the features was studied in the previous experiments, it is important to understand the minimum set of features that can help in classifying the volume. However, it is computationally expensive to use forward feature selection from 6373 features for finding the best set of features that maximize the classification performance. Thus, LASSO feature selection algorithm [17] was used to find the best performing features across all the categories of classification. The LASSO feature selection was implemented, with all the features from OS, using the Feature Selection Library toolbox in MATLAB [18]. Finally, the top 12 features common across all three volume classifications were selected for classification.

### 3.6. Classifier

A linear kernel SVM [19] model was trained on the extracted features, for all three classification tasks. In the dataset, the number of data points in comparison to the number of features was much less. A linear kernel SVM seemed to be a reasonable choice because the cost function of SVM classifiers is less sensitive to outliers in the data and that they can transform data to a high-dimensional space for the classes to be linearly separable, although not linearly separable in low-dimensional space.

### 4. RESULTS & DISCUSSIONS

#### 4.1. Experimental Setup

Features were extracted from the swallowing sounds of 56 subjects whose audio were annotated with swallow components. This helps in obtaining the swallowing sound parameters (baseline features). The OS features were extracted from the swallowing sounds of the same set of subjects. Experiments were performed in a ten-fold cross-validation setup, for which 56 subjects were grouped into ten groups, with almost identical number of subjects in each group. In each fold, seven groups were used for training the classifier, two groups as validation set and the remaining one group was used as the test set. The data in each of the training, validation and test set were normalized feature-wise, using the mean and standard deviation of the respective features from the training set. The dry vs non-dry swallowing classification was categorized into three separate classification tasks namely, dry vs 2ml, dry vs 5ml and dry vs 10ml swallows. The results in the upcoming tables are from an unseen test set.

In the SVM classifier, for optimal hyper parameter (C) selection, grid-search was performed (from $10^{-5}$ to 10, in multiples of 10). The number of data points in dry, 2ml, 5ml and 10ml volumes were not equal. Hence F-score [20] was used as the evaluation metric for the performance of the dry and non-dry swallowing classification system. Statistical tests were performed using Wilcoxon signed rank test for zero median to compare F-score from baseline feature set and OS feature set across folds [21].

#### 4.2. Results of study 1

Table 3 shows F-scores, averaged across all folds, obtained using the baseline and OS features, for all three classification categories, respectively. It was observed that the OS features performed better than the baseline features by 32.88%, 15.67% and 8.35%, for dry vs 2ml, dry vs 5ml and dry vs 10ml, respectively. The p-values indicate that the difference is statically significant compared to baseline features. The OS features showed lower standard deviation across folds compared to baseline features, thus indicating the robustness of the features even with changes in the training data. From the table, it can be observed that the F-score increases with increasing swallow volume using both baseline and OS features. The F-scores, using baseline features showed a relative drop of 46% when the volume of water swallowed changed from 10ml to 2ml, but the drop was only 9% when OS features were used. It is an indication that the OS features were able to discriminate the classes better than the baseline features irrespective of the swallow volume considered.
Table 3: Comparison of mean F-scores (%) of the three categories of classification; values in brackets show the standard deviations

<table>
<thead>
<tr>
<th>Volume / Feature set</th>
<th>Baseline</th>
<th>OS</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dry vs 2ml</td>
<td>37.26 (15.05)</td>
<td>70.14 (8.56)</td>
<td>0.03</td>
</tr>
<tr>
<td>Dry vs 5ml</td>
<td>57.77 (15.05)</td>
<td>73.44 (3.03)</td>
<td>0.0098</td>
</tr>
<tr>
<td>Dry vs 10ml</td>
<td>69.10 (9.67)</td>
<td>77.45 (6.65)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

4.3. Results of study 2

Given that all 6373 features together performed reasonably well in differentiating the volumes, this study was to assess the performance of different groups and subgroups of OS features. The performance of all group A and group B features, and their respective functionals, were separately assessed. From Table 4, it can be seen that, group B features did not yield good accuracy on all three categories of classification. This suggests that glottal excitation features (F0, Jitter etc.) fail to capture swallow sound characteristics. Comparing Table 3 and 4, it is clear that classification performance using Group A features is better than that using all OS features. This could be attributed to its varied feature representation of audio spectrum, which could be useful for classification. Further, experiments were conducted using individual group A features, to find the most discriminating subgroup. Fig. 3 shows the performance using eleven features within group A (other subgroups were omitted because of lower F-scores). It was clear that the performance of all subgroups improved when the volume of water swallowed increased. All the subgroups outperformed the baseline in all classification tasks. Among the group A features, MFCC, RASTA filtered mel-spectrum (AR) and RMS energy emerged consistently as the top performing features for all three categories.

Table 4: Mean F-scores (%) of different Group A features

<table>
<thead>
<tr>
<th>Volume / Class</th>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dry vs 2ml</td>
<td>77.29 (4.19)</td>
<td>58.02 (6.29)</td>
</tr>
<tr>
<td>Dry vs 5ml</td>
<td>78.79 (5.18)</td>
<td>57.60 (4.66)</td>
</tr>
<tr>
<td>Dry vs 10ml</td>
<td>82.03 (7.34)</td>
<td>58.88 (6.04)</td>
</tr>
</tbody>
</table>

4.4. Results of study 3

The LASSO feature selection algorithm was used to rank order features from the acoustic feature set. The feature rank for each fold and the each classification task can be different. Thus, the important features were selected in two ways. In the first one, the top twelve (equal to number of baseline features) features common to all folds, separately for each classification task, were selected. This feature set was referred to as OS-ranked. Next, the top twelve features common across folds and classification tasks were selected. This set was captioned as OS-common. The ranking of both OS-common and OS-ranked features was done using the training set of the respective features. The performance of the OS-ranked and OS-common features is shown in Table 5.

Table 5: Mean F-scores (%) of OS-common and OS-ranked features with standard deviations in ( )

<table>
<thead>
<tr>
<th>Volume / Feature set</th>
<th>OS-common</th>
<th>OS-ranked</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dry vs 2ml</td>
<td>73.55 (6.78)</td>
<td>74.84 (4.85)</td>
<td>0.6523</td>
</tr>
<tr>
<td>Dry vs 5ml</td>
<td>75.88 (9.05)</td>
<td>77.67 (8.17)</td>
<td>0.4316</td>
</tr>
<tr>
<td>Dry vs 10ml</td>
<td>80.68 (4.88)</td>
<td>80.36 (7.61)</td>
<td>1.0</td>
</tr>
</tbody>
</table>

It is clear from the table that, on average, both features performed better than all OS features and the p-values indicate no statistically significant difference between the F-scores obtained using OS-ranked and OS-common features. The OS-common features were found to be different functionals of the smoothed version of Ldns and its derivatives. The spectrum is computed by passing the signal through 26 Mel-frequency bands in the range of 20Hz to 8000Hz, with an auditory weighting with equal loudness curve with cubic-root compression [15]. Ldns is the sum of this spectrum and its derivative is the first order time difference of the feature. The list of functionals that were selected is shown in Table 6.

Table 6: Top 12 OS-common features with Pearson correlation coefficient (PCC) [22] between feature value and the volume of water in [1] (dry was considered as 0ml). Higher PCC indicates better discrimination by respective features of different volumes. $P_k$ indicates the $k$ percentile

<table>
<thead>
<tr>
<th>Ldns</th>
<th>Derivative of Ldns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-quartile range</td>
<td>$(P_{25} - P_{75})$</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>$K = 0.03$</td>
</tr>
<tr>
<td>Minimum segment length</td>
<td>$L_{min} = 0.18$</td>
</tr>
<tr>
<td>Up-level time 25,50,75,90</td>
<td>$T_{up} = 0.03$</td>
</tr>
<tr>
<td>Risetime</td>
<td>$R = 0.023$</td>
</tr>
<tr>
<td>Left curvature time</td>
<td>$L = 0.06$</td>
</tr>
</tbody>
</table>

From the table, it is evident that some robust statistics like percentile were selected. This could be because they are less sensitive to the silence region before/after the swallow sound and, hence, show high positive Pearson correlation coefficient with the volume of water. Functional segment lengths (mean and minimum) measure the number of sub-segments where the Ldns is higher than 25% of its range. Even though the swallow signal boundaries were unknown in the OS features, some of the robust features seem to be indirectly indicating those boundaries. For example, the up-level time functionals (up-level times indicate the number of times the signal crosses a pre-defined threshold [23]), could indicate the length of the SC2 component of their corresponding volumes of water.

5. CONCLUSIONS

In this work, the OS feature set was used to study the acoustic features of swallow sound for classifying bolus volume as dry and nondry classes. The model using the acoustic features performed better than the swallow sound parameters. An improved F-score was observed for three classification tasks considered namely, dry-vs-2ml, dry-vs-5ml and dry-vs-10ml. Across all three volumes of classification, on an average, the 12 functionals of loudness performed better than the OS features. Increasing the size of the dataset and exploring swallow sound-specific features, unlike generic OS features that are designed for speech analysis are future scope to this work of volume classification using swallow sounds.

Acknowledgements: 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.
6. REFERENCES


