healthcare

MDPI

Article

Classifying Dysphagic Swallowing Sounds with Support Vector Machines

Shigeyuki Miyagi [1,*], Syo Sugiyama [1], Keiko Kozawa [2,†], Sueyoshi Moritani [3], Shin-ichi Sakamoto [1] and Osamu Sakai [1]

1 Department of Electronic Systems Engineering, Graduate School of Engineering, The University of Shiga Prefecture, Hikone, Shiga 522-8533, Japan; ot23ssugiyama@ec.usp.ac.jp (S.S.); sakamoto.s@e.usp.ac.jp (S.-i.S.); sakai.o@e.usp.ac.jp (O.S.)

2 Department of Nutrition, School of Human Cultures, The University of Shiga Prefecture, Hikone, Shiga 522-8533, Japan

3 Head, Neck, and Thyroid Surgery, Kusatsu General Hospital, 1660, Yabase, Kusatsu, Shiga 525-8585, Japan; suemoritani@gmail.com

* Correspondence: miyagi.s@usp.ac.jp; Tel.: +81-749-28-9559

† The autor is deceased.

Received: 3 March 2020; Accepted: 16 April 2020; Published: 21 April 2020

Abstract: Swallowing sounds from cervical auscultation include information related to the swallowing function. Several studies have been conducted on the screening tests of dysphagia. The literature shows a significant difference between the characteristics of swallowing sounds obtained from different subjects (e.g., healthy and dysphagic subjects; young and old adults). These studies demonstrate the usefulness of swallowing sounds during dysphagic screening. However, the degree of classification for dysphagia based on swallowing sounds has not been thoroughly studied. In this study, we investigate the use of machine learning for classifying swallowing sounds into various types, such as normal swallowing or mild, moderate, and severe dysphagia. In particular, swallowing sounds were recorded from patients with dysphagia. Support vector machines (SVMs) were trained using some features extracted from the obtained swallowing sounds. Moreover, the accuracy of the classification of swallowing sounds using the trained SVMs was evaluated via cross-validation techniques. In the two-class scenario, wherein the swallowing sounds were divided into two categories (viz. normal and dysphagic subjects), the maximum F-measure was 78.9%. In the four-class scenario, where the swallowing sounds were divided into four categories (viz. normal subject, and mild, moderate, and severe dysphagic subjects), the F-measure values for the classes were 65.6%, 53.1%, 51.1%, and 37.1%, respectively.

Keywords: dysphagia; swallowing sound; machine learning; support vector machine (SVM)

1. Introduction

Dysphagia assessment is a clinically paramount task for detecting dysphagia and presbyphagia in patients, as well as preventing or reducing some diseases caused by a swallowing disorder. For example, pulmonary aspiration due to dysphagia is one of the most common causes of pneumonia, which is the third leading cause of death following malignant neoplasms and heart disease in Japan. Previous research has demonstrated that the death rate caused by pneumonia increases with age [1]. Aspiration pneumonia is the primary type of pneumonia that occurs in older people. Based on clinical research, more than 70% of people with pneumonia appear to have aspiration pneumonia [2].

Currently, the video fluoroscopic (VF) swallowing test is the golden standard for assessing dysphagia [3]. However, the equipment for the VF test is expensive, and the medical attendants performing the test may be exposed to radiation. The video endoscopic (VE) swallowing test is a

common test in clinics and hospitals [4]. Because the equipment for the VE test is small, VE can be performed at the patient's bedside. Nevertheless, the disadvantage of both tests is that they require trained clinicians to perform them. In particular, these tests cannot be performed by nurses, speech-language pathologists, or care helpers.

Alternatively, cervical auscultation is one of the most useful methods for dysphagic assessment [5]. The usefulness of cervical auscultation has been demonstrated throughout the literature. For example, Zenner et al. reported that the results obtained by performing a dysphagia examination with cervical auscultation had a high level of agreement with those obtained from the VF test in patients under long-term care [6]. However, to perform cervical auscultation, trained clinicians or speech-language pathologists are required. Nurses and care helpers in clinics or rehabilitation centers cannot perform cervical auscultation. For overcoming such a situation, an easily operable system for cervical auscultation is expected.

To design such a system, one of the key issues is developing a processing method for the signal obtained from cervical auscultation [7].

In the literature, several researchers have proposed different types of signal processing methods for cervical auscultation. Dudik et al. reviewed several past approaches of signal processing for auscultation [8]. Initially, Takahashi et al. demonstrated the possibility of dysphagic screening using a microphone and accelerometer [9]. In addition, automated classification of abnormal swallowing has been investigated in several studies [10–15]. The relation between swallowing sounds and the mechanism of swallowing [16], as well as the differences in swallowing sounds among various categories of subjects, such as the young, old, normal, or abnormal subjects [17], have been studied previously. Recently, Dudik et al. demonstrated the performance of a deep brief neural network for distinguishing normal swallowing from that in unhealthy patients [18]. The majority of the proposed methods have been applied to two-class problems. However, a conclusive method for automatically classifying swallowing signals based on multiple dysphagic levels has not been realized.

In this study, the support vector machine (SVM)— a machine learning framework— has been applied to swallowing sounds for classifying the subjects into different dysphagic categories. The classification results were compared with the categories based on the dysphagic assessment performed by clinicians. Subsequently, we evaluated the accuracy of the resulting SVM model in classifying the subjects into different dysphagia categories.

2. Method

Healthy subjects, 17 men and 10 women (age range: 21–47, mean: 22.4) with no dysphagic diseases were recruited. Between 2015 and 2017, 78 male patients and 65 female patients (age range: 25–102, mean: 83.3) hospitalized in Kusatsu General Hospital were recruited as dysphagic subjects. The following study protocol received the approval from the ethics committee at the University of Shiga Prefecture and Kusatsu General Hospital.

According to the modified water swallowing test, multiple 3 mL samples of water were given to the subjects. Acoustic sounds were subsequently recorded using a neck-mounted microphone connected to a laptop computer. We used an SH-12iK microphone (Nanzu Musen, Shizuoka, Japan) with a frequency range of 200–3000 Hz. The computer recording software used was Audacity; the sampling rate was set to 8000 Hz, while the recording gain was 0.7. The segment of the swallowing sound was extracted from the successively recorded sound. As the typical swallowing period is known to be about 700 ms in a healthy person [16], a sound segment of a duration of 800 ms was obtained to sufficiently cover most of the swallowing sound. The center of the sound segment corresponded to the position of the peak intensity of the swallowing sound. The total number of segments for the swallowing sound was 170.

Furthermore, clinicians assessed all the healthy subjects and patients according to the VE scoring method proposed by Hyodo et al. [19]. Accordingly, they were categorized into four groups, as listed in Table 1. Category A included healthy subjects. The swallowing sound segments for each subject

were also categorized; these results are also listed in Table 1, along with the number of categorized sound segments.

Table 1. Definition of dysphagia classification categories derived from scoring based on the method proposed by Hyodo et al. [19] using the VE swallowing test.

Category	Total Score Range	Dysphagic Level	Number of Sounds
A	0	normal	104
B	1–4	mild	66
C	5–9	moderate	214
D	10–12	severe	37

2.1. Preprocessing

The preprocessing of the obtained sounds consisted of three steps: noise suppression [20], compensation, and sensitivity of the used microphone. In addition, a filter was applied for the frequency band limit.

In the first step, as the noise profile of the microphone was required for noise suppression, the output sounds obtained from the recording system with no input signals were recorded beforehand in an anechoic chamber. By applying fast Fourier transform (FFT) on the output sounds, the frequency characteristics of the noise was obtained. By subtracting the noise characteristic from the swallowing sound segments in the frequency domain, the silent swallowing sounds were generated.

In the second step, as the sensitivity of the used microphone is required, the amplitude of the recorded signal was measured using the Audacity software. At the beginning, each input had a sinusoidal wave with pressure of 0.1 Pa and a frequency range between 50 Hz–10 kHz in steps of 50 Hz. Fitting a polynomial curve to the variation of the measured amplitude produced the sensitivity level of the microphone, $K_0[f]$. The relative sensitivity, $K[f]$ was derived from the division of $K_0[f]$ using the maximum value of $K_0[f]$. The obtained example of $K[f]$ is demonstrated in Figure 1. By multiplying the inverse of $K[f]$ by the magnitude spectrum of the swallowing sounds, the compensated swallowing sounds were obtained.

Finally, the spectral range of the swallowing sounds was restricted from 200 Hz–3 kHz because of the microphone's acoustic restriction. Furthermore, applying the inverse FFT to the resulting spectrum yielded the final preprocessed swallowing signals. Figure 2 shows an example of the preprocessed swallowing signal.

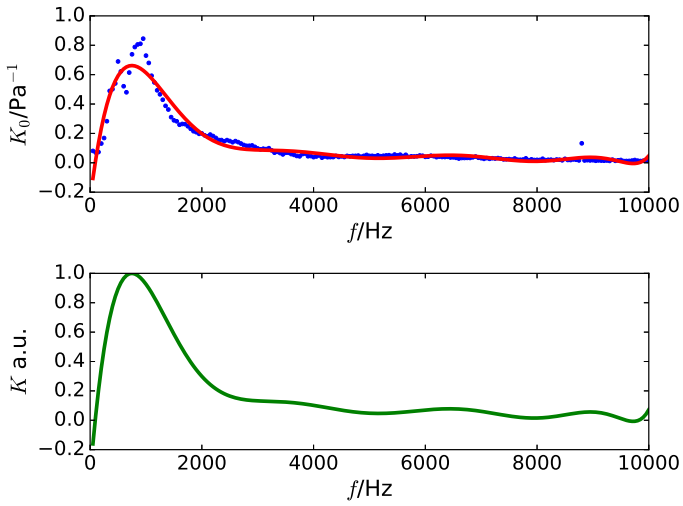


Figure 1. Example of the sensitivity calculation of a microphone.

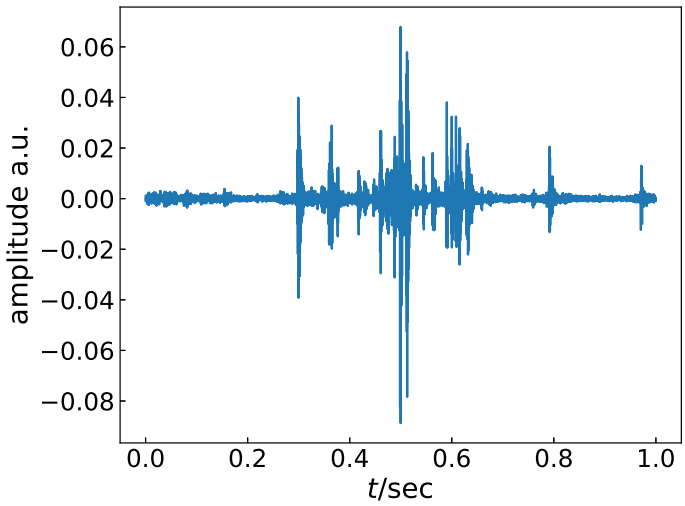


Figure 2. Example of a preprocessed swallowing signal.

2.2. Features

In this subsection, we describe the definitions of the various features that were used for the machine learning process. For extracting the features from the frequency and time-frequency domains of the preprocessed sound signal $x[n]$ $n = 0, 1, 2, \ldots N-1$, we applied the discrete time Fourier series and a short time Fourier transform to $x[n]$. The definitions for both transforms are given as follows:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{j\frac{2\pi k}{N}n} \tag{1}$$

$$X_w[m,k] = \sum_{n=0}^{L-1} x_m[n] e^{j\frac{2\pi k}{L}n}, \tag{2}$$

where $x_m[n]$ denotes the windowed signal by applying the window function $w[n - mS]$ with a window size L, frame number m, and skip width S to $x[n]$. The maximum frame number is defined by $M = N/S$. A spectrogram defined as $|X_w[m,k]|$ from (2) is used in the following Section 2.2.2. An example of the spectrogram is shown in Figure 3a.

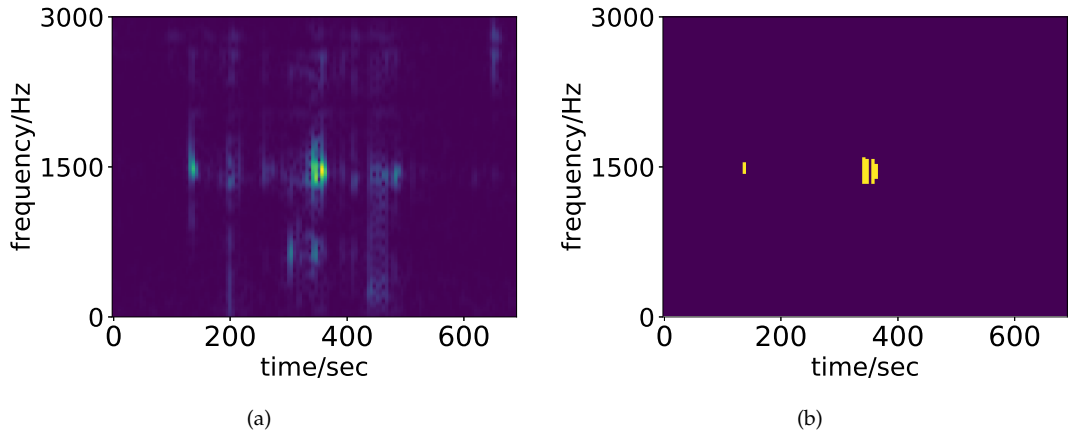


Figure 3. Example of a spectrogram (**a**) and its threshold version (**b**).

2.2.1. Frequency Domain Features

The following conventional features associated with $X[k]$ were used.

- Maximum magnitude of spectra [13] $A_{\max} \equiv \max_{k=0\ldots N/2} |X[k]|$.
- Peak frequency [13] $f_{\max} \equiv \frac{F_s}{N} \arg \max_{k=0,\ldots,N/2} |X[k]|$, where F_s denotes the sampling frequency.
- Frequency average [17]
$$\bar{f} \equiv \sum_{k=0}^{N/2} \frac{f_k\,|X[k]|}{\sum_l^{N/2} |X[l]|}$$
where $f_k = kF_s/N$.
- Standard deviation of frequency
$$\sigma_f \equiv \sqrt{\sum_{k=0}^{N/2} \frac{(f_k - \bar{f})^2\,|X[k]|}{\sum_l^{N/2} |X[l]|}}.$$
- Frequency median and quartile ratio
By considering the magnitude of the spectra $|X[k]|$ as part of a frequency distribution, the quartiles k_1, k_2, k_3 can be calculated, where k_2 corresponds to the median, and the median frequency is given by $Q_2 = k_2F_s/N$. Other indices can also be translated into the frequency $Q_i = k_iF_s/N$. Consequently, the quartile ratio is defined as $Q_{r1} \equiv Q_1/Q_2$, $Q_{r3} \equiv Q_3/Q_2$.
- Total energy
$$E = \frac{1}{N^2} \sum_{k=0}^{N-1} |X[k]|^2$$

2.2.2. Time-Frequency Domain Features

In this study, we propose some features derived from the spectrogram $|X_w[m,k]|$ based on which a binary image $X_{\text{th}}[m,k]$ is obtained via thresholding. These obtained features describe some characteristics of the time-frequency domain signal, including statistical variation, peak locations, and dispersion of peak values.

First, some features associated with the spectrogram in the time-frequency domain are restricted by the indices shown in Figure 4, which are defined as follows.

- Maximum magnitude of the spectrogram $A_w \equiv \max_{m,\,k} |X_w[m,k]|$.
- Peak location of the spectrogram $(m_w,\ k_w) \equiv \arg\max_{m,\,k} |X_w[m,k]|$.
- Relative distances of the peak location from the center of the spectrogram
$$D \equiv \sqrt{D_{\text{H}}^2 + D_{\text{V}}^2}$$
$$D_{\text{H}} \equiv m_w - M/2,\ D_{\text{V}} \equiv k_w - L/4.$$
- Total energy of the spectrogram $E_s \equiv \sum_m \sum_k |X_w[m,k]|^2$.

Furthermore, to describe the time variation of the statistics with respect to the spectrogram, the spectrogram is divided into B-blocks along the m axis, as displayed in Figure 4; subsequently, the following features are defined in each block. The quartile k_{b1}, k_{b2}, k_{b3} for each block is derived from the histogram $H_b[k]$ of the b-th block given by

$$H_b[k] \equiv \sum_{m'=0}^{M/B-1} \left| X_w\left[(M/B)b + m', k\right] \right|.$$

The frequencies for the corresponding quartiles were calculated as $Q_{bi} = k_{bi}F_s/L$. The quartile ratio can be similarly defined by $Q_{rb1} \equiv Q_{b1}/Q_{b2}$, $Q_{rb3} \equiv Q_{b3}/Q_{b2}$. Each block energy is derived from the square sum of the spectrogram, including each block, as

$$E_b = \sum_{m'=0}^{M/B-1} \sum_{k=0}^{L/2} \left| X_w \left[(M/B)b + m', k \right] \right|^2 .$$

In the later experiments, $B = 15$ is used. This means that each block length is approximately 0.05 s under the current sampling rate. The width of the envelope of each peak in the preprocessed signal shown in Figure 2 can be affordably dropped into the block length. Hence, each block holds sufficient information of the difference signal among the other blocks.

Finally, to describe the effects of multiple spectral peaks and the spread of each peak, we propose the following features with respect to the threshold spectrogram $X_{\text{th}}[m,\ k]$ defined as

$$X_{\text{th}}[m,\ k] \equiv \begin{cases} 1 & X_w[m,\ k] \geq \varepsilon_t \\ 0 & \text{otherwise} \end{cases},$$

where $\varepsilon_t \equiv \kappa A_w$ is a threshold level related to the above peak value A_w in the spectrogram and the scale κ. In the following experiments, we used $\kappa = 0.5$. From Figure 3a, there are some local peaks in the spectrogram, except a central main peak. Applying a higher threshold level to the spectrogram may remove the positions of the other spectral peaks. In the case of a lower threshold level, many peak locations, including noisy level spectral peaks, appear. Preliminary experiments determine that $\kappa = 0.5$ could give the appropriate peak location shown in Figure 3b.

- Area of the threshold spectrogram:

$$A_{\text{th}} \equiv \sum_{k=0}^{M-1} \sum_{k=0}^{L/2} X_{\text{th}}[m,\ k]$$

- Ratio of the area of the threshold spectrogram to the complete area of the spectrogram:

$$R_{\text{th}} \equiv A_{\text{th}}/\left(ML/2\right)$$

- Average distance from the center of the spectrogram:

$$\bar{D}_h = \frac{1}{A_{\text{th}}} \sum_{k=0}^{M-1} \sum_{k=0}^{L/2} X_{\text{th}}[m,\ k] \left(m - \frac{M}{2} \right)$$

$$\bar{D}_v = \frac{1}{A_{\text{th}}} \sum_{k=0}^{M-1} \sum_{k=0}^{L/2} X_{\text{th}}[m,\ k] \left(k - \frac{L}{4} \right)$$

$$\bar{D} = \frac{1}{A_{\text{th}}} \sum_{k=0}^{M-1} \sum_{k=0}^{L/2} X_{\text{th}}[m,\ k] \sqrt{\left(m - \frac{M}{2} \right)^2 + \left(k - \frac{L}{4} \right)^2}$$

- Average distance from the peak location of the spectrogram:

$$\bar{D}_{ph} = \frac{1}{A_{\text{th}}} \sum_{k=0}^{M-1} \sum_{k=0}^{L/2} X_{\text{th}}[m,\ k] \left(m - m_s \right)$$

$$\bar{D}_{pv} = \frac{1}{A_{\text{th}}} \sum_{k=0}^{M-1} \sum_{k=0}^{L/2} X_{\text{th}}[m,\ k] \left(k - k_s \right)$$

$$\bar{D}_p = \frac{1}{A_{\text{th}}} \sum_{k=0}^{M-1} \sum_{k=0}^{L/2} X_{\text{th}}[m,\ k] \sqrt{\left(m - m_w \right)^2 + \left(k - k_w \right)^2}$$

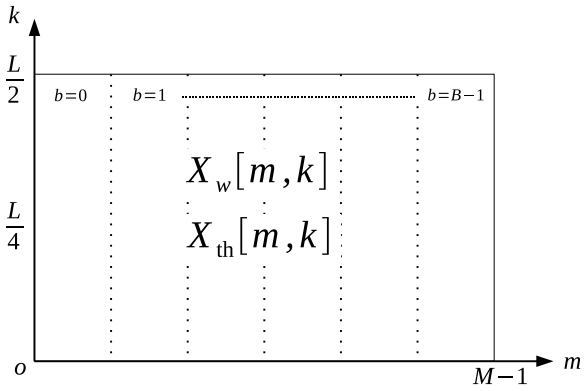


Figure 4. Definition of blocks in the time-space domain.

2.3. Machine Learning

In this subsection, we describe the data set conditions and size adjustments for training the SVMs. Generally, the choice of machine learning method depends on the size of the data set. Deep learning requires vast training data for constructing proper models. For example, the CIFAR-10 data set consists of 60,000 images in 10 classes, with 6000 images per class [21]. It is difficult to collect such a large number of swallowing sounds of patients in a clinical setting. It is experimentally well-known that the SVMs have the advantage of operating with relatively small data sets [22,23]. Therefore, the SVMs were employed in this study. We used the LIBSVM library [24], which is commonly used for implementing SVMs. The efficiency of an SVM with a radial basis function (RBF) kernel depends on the cost and RBF kernel parameter; therefore, these parameters are optimized using the grid-search tool included in the LIBSVM library. The number of swallowing sounds belonging to each category are not always equal to each other. The number of samples from each category used for training should be adjusted to obtain consistent results from the SVM, such that they are almost equal to each other. For example, in the two-class problem, the swallowing sound segments belonging to category A are classified as normal swallowing, and those belonging to categories B, C, and D were classified as abnormal swallowing. Table 1 lists a higher number of sound segments for abnormal swallowing than those for normal swallowing. The number of sound segments that were used was restricted to 104 randomly selected segments. Overall, segments from both groups were divided into 84 training data samples and 20 test data samples. Five different sets of the training and testing data samples were obtained, and all the classified results were averaged to obtain the final classification accuracy. In the four-class problem, the smallest number of the available sound segments was 37 for category D; hence, 37 sound segments were randomly selected from categories A, B, and C. These 37 sound segments were divided into 27 training data samples and 10 test data samples. The remaining process was similar to the two-class problem.

2.4. Choice of Features

When parameter B is set to 15, a total of 83 features can be obtained. These features can create various combinations of features, which require large computational costs for machine learning. For reducing the number of feature combinations for the computational cost, the following restriction, which is a type of filtering method [25], was introduced. The correlation coefficients between the VE scoring and each feature value were calculated beforehand, and they were sorted into the descending order of the absolute value of the correlation coefficients. This variation of the sorted correlation coefficients is shown in Figure 5. Figure 5 shows that the magnitude of the correlation coefficients rapidly decreases, and the decrease is almost linear after the 10th feature. Therefore, the first nine features were used as the candidate features for the machine learning experiments. The classification accuracy for using the top k features of all the 83 features are plotted in Figure 6. The graph is saturated

at approximately four features in the two-class problem, and six features in the four-class problem. Therefore, the feature combinations ${}_9C_4$ and ${}_9C_6$ were studied in these experiments.

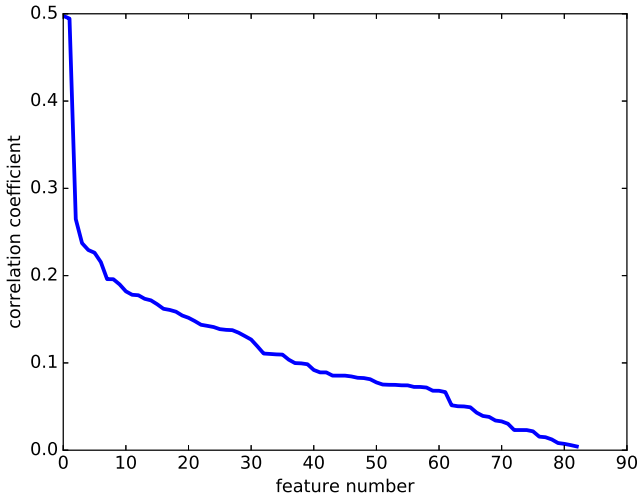
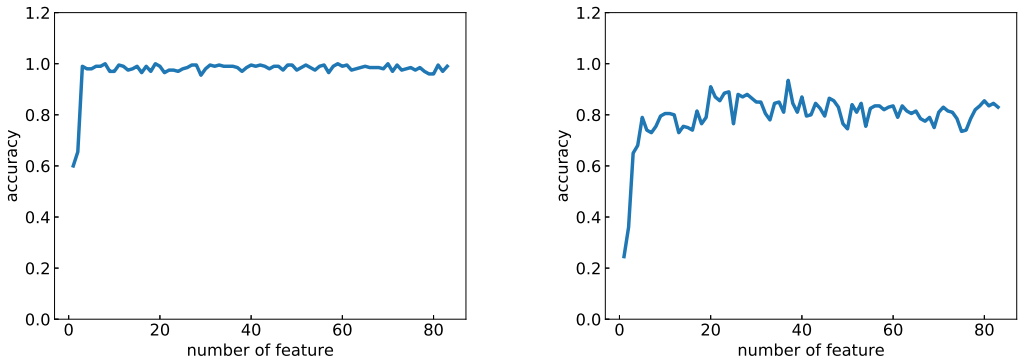


Figure 5. The variation of the correlation coefficients for the total VE scoring performed by the clinicians.



(a) (b)

Figure 6. Variation in classification accuracy by using top k features in (**a**) the two-class problem and (**b**) four-class problem.

3. Results

Table 2 lists the classification accuracy in the two-class problem, where the swallowing sounds of group A are classified as normal, whereas those of groups B, C, and D are classified as abnormal. In this Table, the list has been sorted in descending order of accuracy. The highest value of the accuracy is 0.780, and the highest F-measure is 0.789.

Table 3 summarizes the result of the classification accuracy for the four-class problem. The list is sorted in descending order of total accuracy. The highest total accuracy is 0.460 with features $\bar{D}_v$, $\bar{f}$, A_w, Q_2, $\bar{D}_h$, and Q_{r1}. The F-measures for all feature combinations in each class are plotted in Figure 7. The x-axis and y-axis in those figures indicate the combination number and the F-measure value, respectively.

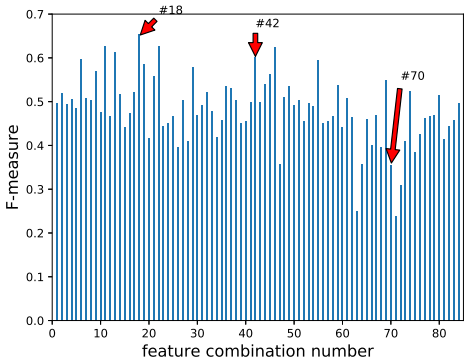
Table 2. Accuracy, precision, recall, and F-measure for each feature combination in the two-class problem. The list is sorted in descending order of accuracy.

Comb. #	Used Features				Accuracy	Precision	Recall	F-Measure
8	$\bar{D}_v$	$\bar{D}$	A_w	Q_2	0.780	0.787	0.790	0.781
18	$\bar{D}_v$	$\bar{D}$	Q_2	$Q_{8,2}$	0.780	0.804	0.750	0.771
54	$\bar{D}_v$	Q_2	$\bar{D}_h$	$Q_{8,2}$	0.780	0.831	0.710	0.760
48	$\bar{D}_v$	$A_{\max}$	Q_2	Q_{r1}	0.775	0.811	0.720	0.762
113	A_w	$A_{\max}$	Q_2	Q_{r1}	0.770	0.763	0.790	0.773
34	$\bar{D}_v$	$\bar{f}$	$\bar{D}_h$	Q_{r1}	0.770	0.752	0.820	0.781
46	$\bar{D}_v$	A_w	Q_{r1}	$Q_{8,2}$	0.770	0.737	0.870	0.789
53	$\bar{D}_v$	Q_2	$\bar{D}_h$	Q_{r1}	0.770	0.750	0.810	0.779
19	$\bar{D}_v$	$\bar{D}$	$\bar{D}_h$	Q_{r1}	0.755	0.717	0.870	0.780
17	$\bar{D}_v$	$\bar{D}$	Q_2	Q_{r1}	0.750	0.808	0.670	0.722

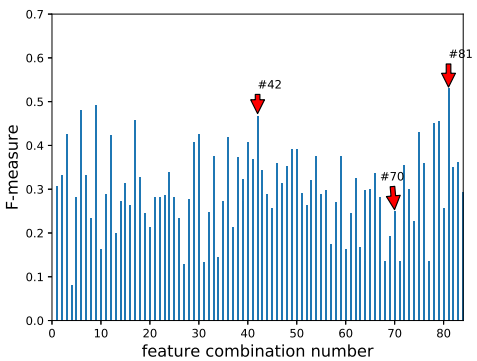
Table 3. Accuracy and F-measures of classes A, B, C, and D for each feature combination in the four-class problem. The list has been sorted in descending order of accuracy.

Comb. #	Used Features					
42	$\bar{D}_v$	$\bar{f}$	A_w	Q_2	$\bar{D}_h$	Q_{r1}
3	$\bar{D}_v$	$\bar{D}$	$\bar{f}$	A_w	$A_{\max}$	Q_{r1}
29	$\bar{D}_v$	$\bar{D}$	A_w	Q_2	Q_{r1}	$Q_{8,2}$
17	$\bar{D}_v$	$\bar{D}$	$\bar{f}$	Q_2	$\bar{D}_h$	Q_{r1}
6	$\bar{D}_v$	$\bar{D}$	$\bar{f}$	A_w	Q_2	Q_{r1}
46	$\bar{D}_v$	$\bar{f}$	$A_{\max}$	Q_2	$\bar{D}_h$	Q_{r1}
9	$\bar{D}_v$	$\bar{D}$	$\bar{f}$	A_w	$\bar{D}_h$	$Q_{8,2}$
45	$\bar{D}_v$	$\bar{f}$	A_w	$\bar{D}_h$	Q_{r1}	$Q_{8,2}$
11	$\bar{D}_v$	$\bar{D}$	$\bar{f}$	$A_{\max}$	Q_2	$\bar{D}_h$
75	$\bar{D}$	A_w	$A_{\max}$	$\bar{D}_h$	Q_{r1}	$Q_{8,2}$

Comb. #	Accuracy	F-Measure (A)	F-Measure (B)	F-Measure(C)	F-Measure(D)
42	0.460	0.602	0.468	0.511	0.145
3	0.420	0.495	0.427	0.500	0.133
29	0.420	0.579	0.407	0.313	0.338
17	0.415	0.521	0.457	0.456	0.095
6	0.410	0.598	0.481	0.279	0.174
46	0.410	0.625	0.360	0.333	0.208
9	0.405	0.571	0.492	0.259	0.141
45	0.400	0.562	0.256	0.400	0.256
11	0.390	0.626	0.288	0.396	0.143
75	0.385	0.385	0.431	0.355	0.356



(a)



(b)

Figure 7. *Cont.*

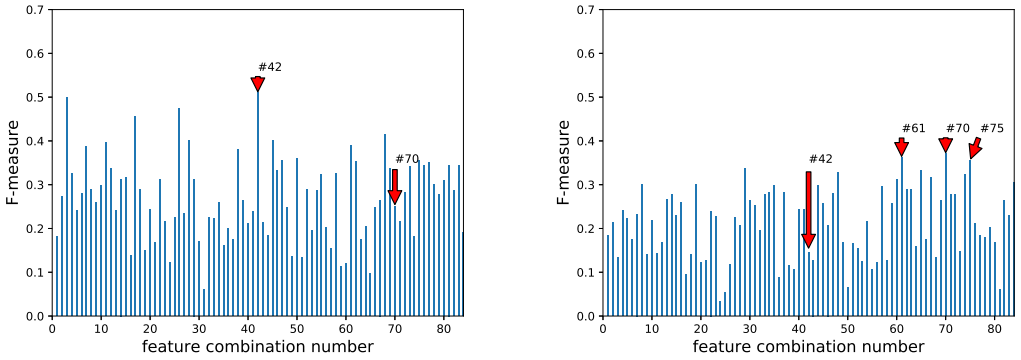


Figure 7. F-measure for each feature combination in classes A, B, C, and D. (**a**) Class A, (**b**) Class B, (**c**) Class C, (**d**) Class D.

4. Discussion

The accuracy in the two-class problem shows similar performance compared to the previous results reported by Mérey et al. [26]. They applied the SVM with an RBF kernel to the dual accelerometry data and demonstrated that the accuracy is 0.806 by using simple feature combination. Based on the feature combination listed in Table 2, the average distance from the center of the spectrogram in the frequency direction, $\bar{D}_v$, as well as in the quartile-related features, such as Q_{r1} and $Q_{8,2}$, are commonly used for obtaining higher accuracy, which suggests that combining these features is useful for increasing the precision of the classification performance. All of these features depend on the position of the frequency peaks or the spectrum bias. As Lee pointed out in [13], the presence of vallecular residue and pyriform sinus residue may change the pharyngeal air space volume, which may cause a change in the acoustic vibration characteristics of the airflow in the pharynx. The feature value of the above features reflects this change.

In the four-class problem, the feature $\bar{D}_v$ and quartile-related features are also included in the combination of the features from Table 3. Although combination #42 yields the highest accuracy and holds a relatively high F-measure for classes A, B, and C, its F-measure in class D is less than 0.15 from Figure 7. Combination #70 indicates the highest F-measure of 0.371 within class D (see Figure 7d). Similarly, combinations #61 and #75 hold the second and third highest F-measures of 0.362 and 0.356, respectively. These feature combinations commonly include $\bar{D}_h$ rather than $\bar{D}_v$, which is included in the feature combinations with high accuracy. These results suggest the following:

1. The characteristic of swallowing sounds for class D is different from that of the other classes.
2. The feature $\bar{D}_h$ relates to the area position with a higher magnitude along the time direction in the spectrograms. Robbins et al. reported that the total duration of oropharyngeal swallowing was significantly longer in the oldest group than in any other younger counterparts [27]. Borr et al., also reported that the duration of the onset time, and deglutition apnea of auscultation for the dysphagic group, was significantly longer than that of the nondysphagic group [28]. These duration variances caused a temporal bias in the spectrogram.
3. When classifying the swallowing sounds into four categories, multiple classifiers or a combination of these might be useful due to the different characteristics of each class.

5. Conclusions

The accuracy of classifying swallowing sounds into two classes and four classes by using SVMs was examined. Although the F-measure reached 78.9% in the two-class scenario, the overall classification accuracy of the four-class scenario was still insufficient when using the classifier as a

stand-alone method for the diagnosis. Conversely, this study revealed that the variation of some feature combinations can serve as useful classifiers for individual categories.

Although these results demonstrate the potential of using a classifier constructed by the SVM, some problems and limitations remain in their use as a practical classifier. The results of the trained SVM relied on the category defined by the total score obtained by Hyodo's VE scoring. The scoring method consisted of four tests. The segmented swallowing sounds may not correspond to the results of the total score of the four tests. To solve this problem, an improved model for considering each test in the scoring should be created. For selecting the features for the SVMs, a filter-based method was used in the experiments and the restriction on the feature combination was introduced. The resulting feature combination may not be optimal. Other types of selection methods, such as wrapper methods or embedded methods, should also be applied and the resulting classifier performances should be compared with each other. It will also be useful to employ other types of machine learning for increasing the classification performance. Deep neural networks is one such option, provided many data sets can be obtained. We are currently working on constructing an improved classifier for considering each test based on Hyodo's VE scoring by using deep neural networks combined with data augmentation. We will report the results of this research in the future.

Author Contributions: Conceptualization, K.K.; Formal analysis, S.S.; Funding acquisition, S.M. (Shigeyuki Miyagi) and O.S.; Investigation, S.M. (Shigeyuki Miyagi), S.S., K.K., S.M. (Sueyoshi Moritani), and S.-i.S.; Methodology, S.M. (Shigeyuki Miyagi); Project administration, S.M. (Shigeyuki Miyagi); Software, S.S.; Supervision, O.S.; Writing—original draft, S.M. (Shigeyuki Miyagi) and S.S.; Writing—review & editing, S.M. (Shigeyuki Miyagi), S.M. (Sueyoshi Moritani), and O.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI Grant Number JP17K01571, Regional ICT Research Center of Human, Industry and Future at the University of Shiga Prefecture, and the Cabinet Office, the Government of Japan.

Acknowledgments: The authors wish to thank all members of the nutrition support team at the Kusatsu General Hospital for helping us with the data acquisition of the swallowing sounds in the examination room.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ministry of Health, Labour and Welfare. General Welfare and Labour. *Ann. Health Lab. Welf. Rep.* **2015**, *1*, 10.
2. Teramoto, S.; Fukuchi, Y.; Sasaki, H.; Sato, K.; Sekizawa,K.; Matsuse, T. High incidence of aspiration pneumonia in community- and hospital-acquired pneumonia in hospitalized patients: A multicenter, prospective study in Japan. *J. Am. Geriat. Soc.* **2008**, *56*, 577–579. [CrossRef] [PubMed]
3. Levine, M.S.; Rubesin, S.E. History and Evolution of the Barium Swallow for Evaluation of the Pharynx and Esophagus. *Dysphagia* **2017**, *32*, 52–72. [CrossRef]
4. Langmore, S.E. History of Fiberoptic Endoscopic Evaluation of Swallowing for Evaluation and Management of Pharyngeal Dysphagia: Changes over the Years. *Dysphagia* **2017**, *32*, 27–38. [CrossRef]
5. Bosma, J.F. Sensorimotor Examination of the Mouth and Pharynx. *Front. Oral Physiol. Physiol. Oral Tiss.* **1976**, *2*, 78–107.
6. Zenner, P.M.; Losinski, D.S.; Mills, R.H. Using cervical auscultation in the clinical dysphagia examination in long-term care. *Dysphagia* **1995**, *10*, 1–64. [CrossRef] [PubMed]
7. Sejdic, E.; Malandraki, G.A.; Coyle, J.L. Computational Deglutition: Using Signal- and Image-Processing Methods to Understand Swallowing and Associated Disorders. *IEEE Sign. Process. Mag.* **2018**, *36*, 138–146. [CrossRef]
8. Dudik, J.M.; Coyle, J.L.; Sejdić, E. Dysphagia Screening: Contributions of Cervical Auscultation Signals and Modern Signal-Processing Techniques. *IEEE Trans. Hum. Mach. Syst.* **2015**, *45*, 465–477. [CrossRef]
9. Takahashi, K.; Groher, M.E.; Michi, K. Methodology for detecting swallowing sounds. *Dysphagia* **1994**, *9*, 54–62. [CrossRef]
10. Hamlet, S.L.; Nelson, R.J.; Patterson, R.L. Interpreting the Sounds of Swallowing: Fluid Flow through the Cricopharyngeus. *Ann. Otol. Rhinol. Larynol.* **1990**, *99*, 749–752 [CrossRef]

11. Das, A.; Reddy, N.P.; Narayanan, J. Hybrid fuzzy logic committee neural networks for recognition of swallow acceleration signals. *Comput. Methods Prog. Biomed.* **2001**, *64*, 87–99. [CrossRef]
12. Lee, J.; Steele, C.M.; Chau, T. Time and time-frequency characterization of dual-axis swallowing accelerometry signals. *Physiol. Meas.* **2008**, *29*, 1105–1120. [CrossRef] [PubMed]
13. Lee, J.; Steele, C.M.; Chau, T. Classification of healthy and abnormal swallows based on accelerometry and nasal airflow signals. *Artif. Intell. Med.* **2011**, *52*, 17–25. [CrossRef] [PubMed]
14. Shirazi, S.S.; Buchel, C.; Daun, R.; Lenton, L.; Moussavi, Z. Detection of swallows with silent aspiration using swallowing and breath sound analysis. *Med. Biol. Eng. Comput.* **2012**, *50*, 1261–1268. [CrossRef] [PubMed]
15. Nikjoo, M.S.; Steele, C.M.; Sejdić, E.; Chau, T. Automatic discrimination between safe and unsafe swallowing using a reputation-based classifier. *BioMed. Eng. OnLine* **2011**, *10*, 100. [CrossRef] [PubMed]
16. Nakayama, H.; Takahashi, K.; Uyama, R.; Hirano, K.; Fukasawa, M.; Nagumo, M. Evaluation of the Sites Where Swallowing Sound Produces and the Acoustic Characteristic of Swallowing Sound in Healthy Adults. *Dent. Med. Res.* **2006**, *26*, 163–174.
17. Dudik, J.M.; Jestrović, I.; Luan, B.; Colyle, J.L.; Sejdić, E. A comparative analysis of swallowing accelerometry and sounds during saliva swallows. *BioMed. Eng. OnLine* **2015**, *14*, 3. [CrossRef]
18. Dudik, J.M.; Coyle, J.L.; El-Jaroudi, A.; Mao, Z.H.; Sun, M.; Sejdić, E. Deep learning for classification of normal swallows in adults. *Neurocomputing* **2018**, *285*, 1–9. [CrossRef] [PubMed]
19. Hyodo, M.; Nishikubo, K.; Hirose, K. New Scoring Proposed for Endoscopic Swallowing Evaluation and Clinical Significance. *Nippon Jibiinkoka Gakkai Kaiho* **2010**, *113*, 670–678. [CrossRef]
20. Vaseghi, S.V. Chap-12 Signal enhancement via spectral amplitude estimation. In *Advanced Digital Signal Processing and Noise Reduction*; Wiley: West Sussex, UK, 2008; pp. 321–339.
21. Krizhevsky, A. Learning Multiple Layers of Feature from Tiny Images. Master's Thesis, Department of Computer Science, University of Toronto, Toronto, ON, Canada, 2009.
22. Raschka, S. Available online: https://www.kdnuggets.com/2016/04/deep-learning-vs-svm-random-forest.html (accessed on 3 April 2020).
23. Amatriain, X. Available online: https://www.quora.com/What-are-the-advantages-of-different-classification-algorithms (accessed on 3 April 2020).
24. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
25. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28 [CrossRef]
26. Mérey, C.; Kushki, A.; Sejdić, E.; Berall, G.; Chau, T. Quantitative classification of pediatric swallowing through accelerometry. *J. NeuroEng. Rehabil.* **2012**, *9*, 1–8.
27. Robbins, J.; Hamilton, J.W.; Lof, G.L.; Kempster, G.B. Oropharyngeal swallowing in normal adults of different ages. *Gastroenterology* **1992**, *103*, 823–829. [CrossRef]
28. Borr, C.; Hielscher-Fastabend, M.; Lucking, A. Reliability and validity of cervical auscultation. *Dysphagia* **2007**, *22*, 225–234. [CrossRef] [PubMed]