CLARKSON UNIVERSITY

Automatic method of acoustical swallowing detection for monitoring of ingestive behavior

A Dissertation

By

Oleksandr Makeyev

Coulter School of Engineering

Submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy, Engineering Science

April 13, 2010

Accepted by the Graduate School

_____, _____

Date, Dean of the Graduate School

The undersigned have examined the dissertation entitled '**Automatic method of acoustical swallowing detection for monitoring of ingestive behavior**' presented by **Oleksandr Makeyev**, a candidate for the degree of **Doctor of Philosophy (Engineering Science)** and hereby certify that it is worthy of acceptance.

Date

Prof. Edward Sazonov

Prof. Stephanie Schuckers

Prof. Charles Robinson

Prof. William Wilcox

Prof. Janice Searleman

Abstract

Our understanding of the etiology of obesity and overweight is incomplete due to a lack of objective and accurate methods for monitoring ingestive behavior in the free living population. As a step toward such monitoring, an objective and automatic approach to detect periods of food intake based on data from non-invasive and wearable swallowing sensor is proposed. Our research has shown that the frequency of swallowing may be used to detect food intake. Therefore, an automatic swallowing detection methodology was proposed to produce time series of automatically detected swallows, which can further be used to detect and characterize food intake. This methodology should be suitable for obese subjects and able to separate swallowing sounds from sounds that originate from respiration, intrinsic speech, head movements, food intake, and ambient noise. In this dissertation the feasibility of artifact elimination with an acoustical swallowing detection method is described. An automatic swallowing detection methodology that meets the aforementioned requirements is proposed. It was tested on a large database collected from individuals with various degrees of adiposity during periods of food intake and resting, in conditions resembling free-living. Automatically detected swallows are further used to detect periods of food intake in both personalized and non-personalized models. These models can be directly implemented in a wearable device. Such a device can potentially be used in free-living conditions improving our understanding of eating behaviors associated with obesity and other eating disorders and providing the real-time biofeedback to individuals. To the best of our knowledge, this is the first attempt of fully automatic detection of food intake based on data from a wearable non-invasive swallowing sensor. Experimental results suggest efficiency and reliability of

the proposed automatic swallowing detection methodology and its potential for monitoring of ingestive behavior based on swallowing.

Acknowledgements

I'd like to take this portion of my dissertation to state my sheer gratitude to those who have made this all possible.

First and foremost, I would like to thank my Mom for everything she has done for me. With all my heart I dedicate this thesis to her; it is the very least I can do fully express the extent of my gratitude.

I thank all my family for always being there for me and to Dr. Ernst Kussul for being my mentor.

I thank Prof. Edward Sazonov and Prof. Stephanie Schuckers for guiding me through this work and believing in me. I thank all Clarkson professors I worked with and especially my committee members for setting me an example to aim for both as a professional and as a human being. I thank all the lab mates during these years; it's been a pleasure to work with you guys. I especially thank Paulo Lopez-Meyer and I hope to repay all the favors I owe him one day.

I thank all my weird ingenious friends and all the kendo people I've met in Ukraine, Mexico and US for making me feel at home wherever I go.

Last but not the least, I thank Anna for being supportive, understanding and overall making my life better.

I appreciate you all very much.

Table of Contents

Chapter Page

Abstract iii

Acknowledgements v

Table of Contents vi

List of Tables viii

List of Figures ix

1. Introduction 1

2. Swallowing Detection Background 6

3. Feasibility of Artifact Elimination with an Acoustical Swallowing Detection Method 12

3.1. Data Collection and Preprocessing 13
3.2. LIRA Neural Classifier 16
3.3. Experimental Results 18
3.4. Discussion 20
3.5. Conclusion 22

4. Human Study and Data Collection Process 24

4.1. Sensors 26
4.2. Software 30
4.3. Data collection protocol 32
4.4. Conclusions 34

5. Automatic Detection of Swallowing Events by Acoustical Means 35

5.1. Abstract 35
5.2. Introduction 36
5.3. Acoustical detection of swallowing events 39
5.4. Data collection 42
5.5. Methodology 43
5.5.1. Feature Extraction by Wavelet Packet Decomposition 43
5.5.2. Feature Extraction by Mel-Scale Fourier Transform 44
5.5.3. Support Vector Machines 45
5.5.4. Selection of Optimal Epoch Duration and Decomposition Level 46

5.5.5. Training and Validation 46
5.5.6. Accuracy of Detecting Swallowing Instances 47
5.6. Results 48
5.7. Discussion 50
5.8. Conclusion 54

6. Automatic Food Intake Detection Based on Swallowing Sounds 56

6.1. Abstract 56
6.2. Introduction 57
6.3. Related Work 60
6.4. Methodology 65
6.4.1. Human Study 65
6.4.2. Automatic Detection of Swallowing Instances 67
6.4.3. Automatic Detection of Food Intake 70
6.5. Experimental results 71
6.5.1. Intra-subject Food Intake Detection Models 71
6.5.2. Inter-subject Food Intake Detection Models 74
6.6. Discussion 77
6.7. Conclusion 81

7. Overall Conclusions and Future Work 83

7.1. Feasibility of Artifact Elimination with an Acoustical Swallowing Detection Method 83
7.2. Automatic Detection of Swallowing Events by Acoustical Means 83
7.3. Automatic Food Intake Detection Based on Swallowing Sounds 85
7.4. Future Work 86

References 91

Appendix A 100

Appendix B 112

Appendix C 116

List of Tables

Table Page

Table 1: Comparison of recognition rates for combination of LIRA with CWT and STFT 19
Table 2: Accuracy obtained in swallowing detection for three-fold cross-validation. 49
Table 3: Effects of preprocessing (PCA) and postprocessing (smoothing algorithm) on average accuracy for intra-subject model 73
Table 4: Effects of preprocessing (PCA) and postprocessing (smoothing algorithm) on average accuracy for inter-subject model 76

List of Figures

Figure Page

Figure 1: Examples of spectrograms of (columns): a) swallowing sounds, b) talking, c) head movements, d) segments of music recordings..... 15
Figure 2: Examples of scalograms of (columns): a) swallowing sounds, b) talking, c) head movements, d) outlier sounds..... 16
Figure 3: Structure of the LIRA neural classifier. 16
Figure 4: Suggested sensor locations. 25
Figure 5: a) IASUS NT throat microphone, b) time series and spectrogram of consuming 5 peanuts with 3 swallows, c) time series and spectrogram of a gulp of water of arbitrary size. 27
Figure 6: Block diagram of the multi-modal data collection system. 29
Figure 7: Scoring software graphical user interface: 1) activity mark, 2) bites/chews/swallows track, 3) ambient sound signal, 4) throat microphone signal, 5) bone conduction microphone signal, 6) strain sensor signal, 7) user button signal..... 31
Figure 8: Feature extraction: a) A 4.0s fragment of a sound recording including a swallow, b) features extracted by WPD processing, c) features extracted by msFS processing. Frequencies are shown for the center of the extracted band. 45
Figure 9: Examples of: a) true positive, b) false positive, c) true negative, d) false negative. Each number represents a class label for an epoch ('-1' – non-swallow epoch, '1' – swallow epoch)..... 48
Figure 10: Accuracy of swallowing sound recognition as a function of epoch duration and decomposition level: a) msFS with no lags, b) WPD with no lags, c) msFS with K=1, (3 lags), d) WPD with 3 lags. 49
Figure 11: Distribution of average weighted accuracy in classification of epochs and swallowing events versus subject's BMI and corresponding linear fit of the data..... 50
Figure 12: Scheme of the two-stage automatic food intake detection. 59
Figure 13: ROC curves for intra-subject model: per subject and average for all subjects. 73
Figure 14: Distribution of average accuracy in per-epoch and per-swallow swallowing detection versus subject's BMI for intra-subject model. 74
Figure 15: ROC curves for inter-subject model: per subject and average for all subjects. 76
Figure 16: Distribution of average accuracy in per-epoch and per-swallow swallowing detection versus subject's BMI for inter-subject model. 77
Figure 17: Examples of positioning of the throat microphone for four data collection visits of the same subject. 113
Figure 18: Example of the incorrect positioning of the throat microphone..... 113
Figure 19: Examples of inefficiency of the proposed band in cases of subjects with: a) very low BMI, b) very high BMI. 114

Figure 20: Example of inefficiency of the proposed band in holding the throat microphone in place during the same visit: a) beginning of the first session, b) end of the third session. 115

1. Introduction

The obesity epidemic has been suggested as a primary cause for a potential decline in US life expectancy [Olshansky et al. 2005]. In 2005, over 40% of Americans were obese with Body Mass Index (BMI) of more than 30 and over 73% were overweight with BMI of more than 25 [WHO 2006]. Obesity contributes to an increased risk of heart disease, hypertension, diabetes, and some cancers [Carmelli et al. 1997, James 1998] and is now considered a risk factor for cardiovascular disease [Eckel and Krauss 1998]. Millions of Americans are attempting to lose weight at any time, but the rate of success at preventing weight regain is only 5-20% [Wyatt and Hill 2002].

The modern lifestyle is the most probable cause of obesity epidemic. We are surrounded by inexpensive, available, highly palatable and highly caloric food. At the same time the level of physical activity is significantly reduced in comparison to just several decades ago. Recent research indicates that control of food intake may be the primary factor for maintaining a healthy weight in such an environment [Flatt 1996].

At the present time, there is no accurate, inexpensive, non-intrusive way to objectively monitor food intake behavior in free living conditions.

Several methods have been proposed to measure food intake including observation, weighed food records, estimated records, diet history, food-frequency questionnaires, food recall methods, and others. Indirect measurement of food intake through the use of doubly-labeled water [Schoeller 1988] has been used as a gold standard to assess other methods for measuring energy intake [Livingstone and Black 2003]. In a review of 43

studies comparing these methods to doubly-labeled water, the majority suffers from underestimation of energy intake on the order of 0.84 (ratio of estimate to actual intake) [Livingstone and Black 2003]. Observation gives the best agreement, but is expensive and may not be representative of the typical behavior of the subjects. All of the methods based on self-reporting have significant under-estimation [Mertz et al. 1991, Subar et al. 2003, Prentice et al. 1989, Weber et al. 2001, Champagne et al. 2002] due to many factors, including under-reporting, under-eating, recording burden, psychological and behavioral aspects.

While methods utilizing doubly-labeled water are considered accurate and can be used for free-living individuals, they are expensive, difficult to use for large studies, and provide an integrated measure of energy intake across subjects and across a period of weeks that does not allow measurement of daily, individual energy intake patterns. On the other hand, individual methods may provide a good estimate of portions of energy intake, but lack information on the specific patterns of food intake throughout the day. For example, one study showed that meal intake is reported well in contrast to snacks [Poppitt et al. 1998], while others show other differences. In summary, comparison of food intake measures, particularly those designed for free living individuals, suffer from significant under-estimation which is related to the characteristics of the subjects themselves and also to the patterns of the food they ingest.

People who would like to monitor and possibly lose weight would love to have a device that can objectively monitor their ingestive behavior over time and use this device for behavioral modification programs. Our long term goal is the development of an

affordable wearable device to non-invasively detect instances of swallowing (deglutition) in free living individuals as a way to objectively determine when and how often food consumption is taking place. Focusing on monitoring of ingestive behavior, an important task is to identify bouts of food-related swallowing as a means of quantifying the number and length of eating periods per day. The target device should not require special fitting of the sensors and should not exceed the capabilities of a typical low-power embedded processor. Such a device would be suitable for free living individuals who are not required to cooperate in the reporting of food consumption beyond the wearing of the device and may be used for both advising individuals about moderating their food intake in the same way pedometers are used for behavioral modifications related to energy expenditure and in research and clinical applications to study behavioral patterns of food consumption related to obesity and other eating disorders.

Specifically, such a device is envisioned to consist of two components: a wireless component including a microphone and wireless transmitter will capture the swallowing sounds and send the sound signals to a second component, a portable device, such as a cell phone, iPod, or PDA. The second component will process the sounds, distinguish between swallows/non-swallows, and determine bouts of eating.

Our recent research [Sazonov et al. 2009b] has shown that frequency of swallowing can serve as a predictor for accurate detection of food intake, differentiation between liquid and solid foods and estimation of ingested mass, with high frequency of swallowing being indicative of ingestion. Therefore, to create the envisioned device, there is a need for an automatic swallowing detection methodology that would be suitable for obese

subjects and would be able to separate swallowing sounds from sound artifacts that originate in talking, head movements, food ingestion, respiration, etc. With such methodology algorithms presented in [Sazonov et al. 2009b] can be applied as the second step of processing to detect and characterize food intake from the time series of automatically detected swallows.

In this dissertation the feasibility of artifact elimination with an acoustical swallowing detection method is shown and an automatic swallowing detection method is proposed and tested on a large database collected from 20 individuals with various degrees of adiposity during periods of food intake and resting in conditions resembling the free-living. Automatically detected swallows are further used to detect periods of food intake in both personalized (intra-subject) and non-personalized (inter-subject) models that can be directly implemented in the envisioned device. Experimental results suggest efficiency and reliability of the proposed methodology and its potential for monitoring of ingestive behavior based on swallowing.

This dissertation is organized as follows: In chapter 2 the background on assessment of swallowing sound signals and currently used swallowing detection methods is presented. In chapter 3 the feasibility of acoustical artifact elimination with a swallowing detection method is shown. Description of the human study and collection of data that was used to validate automatic swallowing and food intake detection methodologies is presented in chapter 4. Chapters 5 and 6 contain more specific background, description and testing results obtained for automatic swallowing and food intake detection methodologies

correspondingly. Overall conclusions summarizing the unique contributions of this dissertation and an overview of future work are presented in chapter 7.

2. Swallowing Detection Background

At present, various non-invasive methods are proposed for swallowing assessment based on digital signal processing techniques [Das et al. 2000, Lazareck and Moussavi 2002, Lazareck and Moussavi 2004, Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2005, Aboofazeli and Moussavi 2006]. Some of these works concentrate on differentiating between individuals with and without swallowing dysfunction or dysphagia [Das et al. 2000, Lazareck and Moussavi 2004, Aboofazeli and Moussavi 2005] while others are focused on automated decomposition of tracheal sounds into swallowing and respiratory segments [Lazareck and Moussavi 2002, Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2006].

In [Das et al. 2000] two sets of hybrid fuzzy logic committee neural networks (FCN) were proposed for recognition of dysphagic swallows, normal swallows and artifacts (speech, head movement) and tested on data obtained from two groups of 12 normal and 16 dysphagic subjects. The subjects in the upright position were administered small quantities of food as determined by the clinician and were asked to swallow on command. Five features (number of zero crossings, average power, average frequency, maximum power, and frequency at maximum power) were used as inputs of the FCN. Recognition rate of 97.1% was obtained.

In [Lazareck and Moussavi 2004] and [Aboofazeli and Moussavi 2005] two methods of classification of normal and dysphagic swallows are proposed. Both methods are tested on the data obtained from two groups of 15 normal and 11 dysphagic subjects. Subjects were fed three textures: “semisolid”, “thick liquid”, and “thin liquid”. A total of 350

swallowing signals were utilized. In [Lazareck and Moussavi 2004] eleven features, including waveform dimension, magnitude, and average power calculated for different frequency bands were used. Discriminant analysis was performed for classification of normal and dysphagic swallowing sound signals for each texture. In [Aboofazeli and Moussavi 2005] four features including optimum time delay and correlation dimension of the opening and transmission phases were used as input to 3-nearest neighbor classifier. In both studies, the subject was classified as dysphagic if more than 50% of his/her swallows were classified as dysphagic and not dysphagic otherwise. Both methods [Lazareck and Moussavi 2004] and [Aboofazeli and Moussavi 2005] were able to classify swallows correctly in 24 out of 26 subjects.

The disadvantage of [Das et al. 2000, Lazareck and Moussavi 2004, Aboofazeli and Moussavi 2005] is time-consuming and subjective manual extraction of swallowing signals from the recordings through repeated listening and monitoring of the signal in the time and frequency domains. Several methods for automatic detection of swallowing signals in the tracheal sound signal have been proposed [Lazareck and Moussavi 2002, Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2006].

In [Lazareck and Moussavi 2002] utilization of three tracheal sound features including autoregressive coefficients, root-mean-square values of the signal in time domain, and the average power of the signal within different frequency bands was proposed. The method was tested on six respiratory and swallowing sound recordings obtained from healthy subjects. Three features were used for preliminary classification based on cluster analysis and 95% confidence interval thresholds. Definitive classification was performed by a

"smart" algorithm simulating the trained physician tracking patterns within the sound signal. The recognition rate of 78.54% was obtained.

In [Aboofazeli and Moussavi 2004] a method based on multilayer feed forward neural networks was proposed. The method was tested on 18 tracheal sound recordings (with an average length of 40 seconds) for 7 healthy subjects. During the data collection process all participants were fed thin liquid in separate boluses of fixed size resulting in a total of 253 swallows. Root-mean-square values, waveform fractal dimension, and average power of the signal over 150-450Hz were calculated for each signal segment, 5 preceding segments and 5 following segments, along with the mean values for each feature for all 11 segments were used as input to the multilayer feed forward neural network with one hidden layer of 9 neurons. The average rate of missed swallows and false swallow detections were of 8.3% and 9.5% respectively.

In [Aboofazeli and Moussavi 2006] a discrete wavelet transform based filter with iterative sequences of multiresolution decomposition and reconstruction is proposed. The data for this study were adopted from [Lazareck and Moussavi 2004]. An average recognition rate of 93% was obtained with an average rate of missed swallows and false swallow detections of 4% and 3% respectively.

These methods [Lazareck and Moussavi 2002, Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2006] have limitations.

First, all of the aforementioned methods are based on utilization of accelerometers placed over the suprasternal notch of the trachea for data collection. While direct application of

accelerometers to detect swallows in obese people was not evaluated there is some indirect indication that sensors detecting larynx and laryngeal prominence movements may not be reliable for obese people. For example, in [Pehlivan et al. 1996] a piezoelectric strain sensor held in place between the cricoid and the thyroid cartilages by a band of elastic material is proposed for measurement of frequency of spontaneous swallowing. The sensor detected upward and downward motion of the larynx produced by swallowing. Reported data indicate that a laryngeal strain sensor is not appropriate for obese subjects, since under chin fat pads inhibit reliable detection of swallows. Furthermore, in [Lear et al. 1965] the authors reported failure of a pneumatic method to detect swallowing in subjects where "a mass of soft tissue overlay the laryngeal prominence and masked the surface disturbance caused by its movements". They also reported successful experimentation with an acoustical method that detected "a short sharp noise … on the skin lateral to the laryngeal prominence" and pointed out that "when detected by instruments of suitable sensitivity, the swallowing sound, regardless of its intensity, can be readily distinguished from other noises heard in the area, such as intrinsic sounds of speech, belching, coughing and snoring, and the extrinsic sounds generated by movements of clothes, sheets, etc. against the recording device" which suggests reliability of sound sensors for obese people.

Second, all of the proposed methods for automated detection of swallowing [Lazareck and Moussavi 2002, Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2006] have only assessed separation of swallowing segments from respiratory segments and rely heavily on characteristics of breath sounds. For example, the methodology that yielded the best average recognition rate of 93% [Aboofazeli and Moussavi 2006] is

based on assumption that “as swallowing sounds have more non-stationarity compared with breath sounds, they have larger components in many wavelet scales whereas wavelet coefficients of breath sounds in higher wavelet scales are small” suggesting potential vulnerability of this method to non-stationary artifacts. In practical situations artifacts such as talking, throat clearing, head movements, etc., may be confused with swallowing and breath, decreasing the efficiency of the recognition [Das et al. 2000]. In [Lazareck and Moussavi 2002] the authors state that “all the swallowing ‘click’ sounds were classified correctly but some segments within the swallowing sections as well as some segments in forceful expiration sections were misclassified (especially for segments that were neither breathing nor swallowing but included a noise due to tongue movement)” indicating vulnerability of the methodology to artifacts. Furthermore, these studies did not take into account sound artifacts that originate from food ingestion (bites, chewing, etc.) elimination of which is very important for monitoring of the ingestive behavior under free living conditions.

Finally, the description of all the proposed automated swallowing detection methods [Lazareck and Moussavi 2002, Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2006] lacks a clear indication on whether the recognition rates are reported for group or individual models and how the dataset was divided for training and validation which is very important for evaluation of the suitability of the method for creation of a monitoring device. Based on the given information one can assume that all the results are reported for group models using leave-one-out cross-validation with the recognition rate averaged across subjects.

Therefore, it can be seen that there is a strong need for an automatic swallowing detection method that would be suitable for obese subjects and would be able to separate swallowing sounds from sound artifacts that originate in talking, head movements, food ingestion, and respiration. In order to allow creation of a wearable monitoring device, this method should be noninvasive, not require special fitting of the sensors and it should not exceed the capabilities of a typical low-power embedded processor.

3. Feasibility of Artifact Elimination with an Acoustical Swallowing Detection Method

Related publications:

- Makeyev O, Sazonov E, Schuckers S, Melanson E and Neuman M (2007a) "Limited receptive area neural classifier for recognition of swallowing sounds using short-time Fourier transform" Proc. Int. Joint Conf. on Neural Networks IJCNN'2007 (Orlando, USA) 1417.1-6.
- Makeyev O, Sazonov E, Schuckers S, Lopez-Meyer P, Melanson E and Neuman M (2007b) "Limited receptive area neural classifier for recognition of swallowing sounds using continuous wavelet transform" Proc. of 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society EMBC'2007 (Lyon, France) 3128-31.
- Makeyev O, Sazonov E, Schuckers S, Lopez-Meyer P, Baidyk T, Melanson E and Neuman M (2008b) "Recognition of swallowing sounds using time-frequency decomposition and limited receptive area neural classifier" Proc. of 28th SGAI Int. Conf. on Innovative Techniques and Applications of Artificial Intelligence (Cambridge, UK) 33-46.

In this chapter the ability to recognize swallow signals and eliminate artifacts with high accuracy using a novel acoustical swallowing sound recognition technique combining the limited receptive area (LIRA) neural classifier with two time-frequency decomposition methods, short-time Fourier transform (STFT) and continuous wavelet transform (CWT) is demonstrated [Makeyev et al. 2007a, Makeyev et al. 2007b, Makeyev et al. 2008b].

The novelty of the proposed technique consists of the following: time-frequency decomposition methods commonly used in sound recognition increase dimensionality of the signal and require steps of feature selection and extraction. Usually feature selection is based on a set of empirically chosen statistics, making the pattern recognition dependent on the intuition and skills of the investigator. A limited set of extracted features is then presented to a classifier. The proposed method avoids the steps of feature

selection and extraction by delegating them to a LIRA neural classifier that utilizes the increase in dimensionality of the signal to create a large number of random features in the time-frequency domain that assure a good description of the signal without prior assumptions of the signal properties. Features that do not provide useful information for separation of classes do not obtain significant weights during classifier training.

The LIRA neural classifier was developed as a multipurpose image recognition system [Kussul et al. 2006] and tested with promising results in different image recognition tasks including: handwritten digit image recognition [Kussul and Baidyk 2004], micro device assembly [Baidyk et al. 2004], mechanically treated metal surface texture recognition [Makeyev et al. 2008a], face recognition [Kussul et al. 2004], micro work piece shape recognition [Kussul et al. 2006] and recognition of postural allocations [Sazonov et al. 2007]. Application of the LIRA-based image recognition technique to a two-dimensional power spectrum, such as a spectrogram in case of a short-time Fourier transform (STFT) or a scalogram in case of continuous wavelet transforms (CWT) is presented below.

3.1. Data Collection and Preprocessing

A commercially available miniature throat microphone IASUS NT (IASUS Concepts Ltd.) located over the laryngopharynx was used during the data collection process. Throat microphones convert vibration signals from the surface of the skin rather than pick up waves of sound pressure, thus reducing the ambient noise. Throat microphones also pick up such artifacts as head movements and talking that should not be confused with swallowing sounds.

Twenty sound instances were recorded for each of three classes of sounds (swallow, talking, head movement) for a healthy subject without any history of swallowing disorders, eating or nutrition problems, or lower respiratory tract infection. To record the swallowing sound the subject was asked to consume water in boluses of arbitrary size. For head movement artifact recording the subject was asked to turn his head to a side and back. To record the speech artifact the subject was asked to say the word “Hello”. Sound signals for each class were amplified and recorded with a sampling rate of 44100 Hz.

A fourth class of outlier sounds that consisted of random segments of music recordings was introduced to demonstrate the ability of the neural classifier to reject sounds with weak intra-class similarity and no similarity with the other three classes.

Next, swallowing, head movement, and talking sounds were extracted from the recordings in segments of 65536 samples (approximately 1.5 s) each using the following empirical algorithm: the beginning and end of each sound were found using a threshold set above the background noise level; then the center of mass was calculated for each sound and used to center the corresponding sound instance in the recognition window. Spectrograms of each segment were calculated with a window of 512 samples extracted using a Hanning window function and processed by STFT with 50% window overlap. Due to the limited signal bandwidth higher frequencies do not contain significant energy of the original time domain signal and can be eliminated from the spectrogram. Truncating the spectrogram from 512x256 pixels to 256x256 pixels preserves most of the signal energy and eliminates insignificant harmonics. Scalograms of each segment were calculated with a Morlet mother wavelet with wavenumber of 6, 7 octaves and 16

suboctaves. To compare pattern recognition accuracy on time-frequency decompositions produced by CWT and STFT the following processing was applied to the scalograms: a mirror image of the scalograms across the abscissa was created and combined with the original; the resulting image was resized to 256x256 pixels using bicubic interpolation.

Eighty grayscale spectrogram images (20 for each of 4 classes) comprise the first image database that was used in training and validation and eighty grayscale scalogram images (20 for each of 4 classes) comprised the second image database. Examples of spectrogram and scalogram images are presented in Fig. 1 and Fig. 2 allowing the direct visual comparison to be drawn.

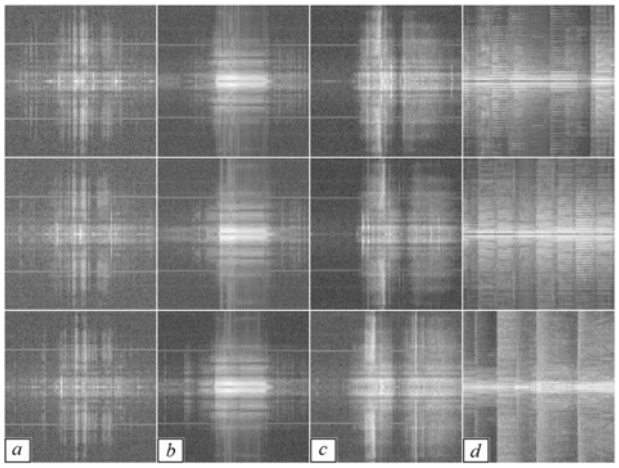


Figure 1: Examples of spectrograms of (columns): a) swallowing sounds, b) talking, c) head movements, d) segments of music recordings.

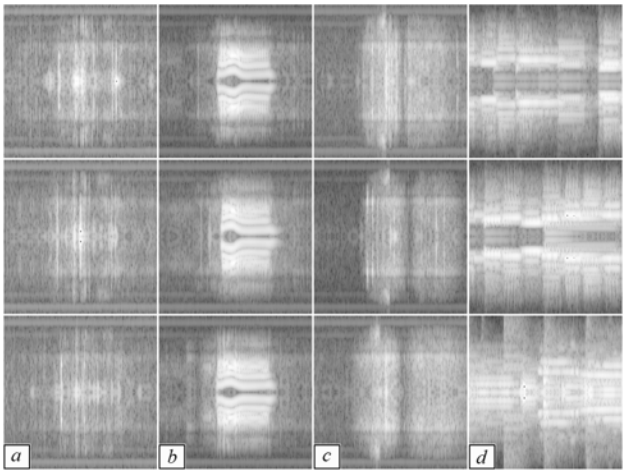


Figure 2: Examples of scalograms of (columns): a) swallowing sounds, b) talking, c) head movements, d) outlier sounds.

3.2. LIRA Neural Classifier

The LIRA neural classifier is a multi-layer perceptron that consists of *S*-layer, *I*-layer, *A*-layer and *R*-layer (Fig. 3).

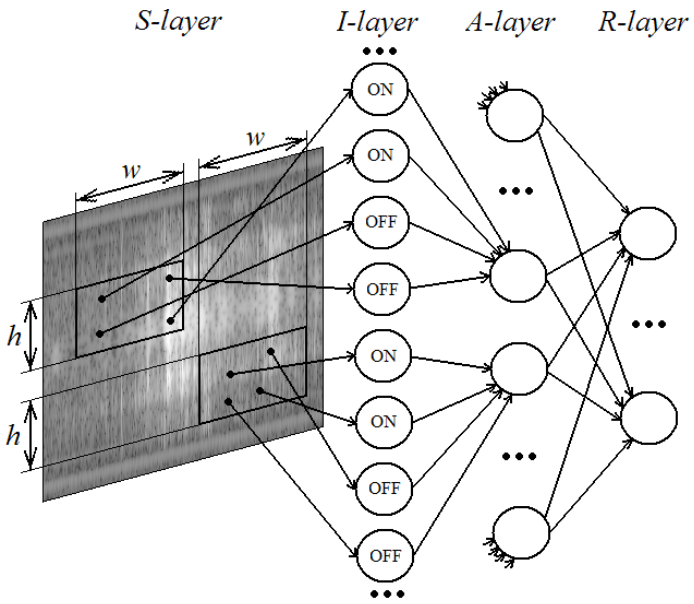


Figure 3: Structure of the LIRA neural classifier.

Sensor *S*-layer corresponds to the input image. Associative *A*-layer is connected to the *S*-layer through the intermediate *I*-layer with randomly selected non-trainable connections. The set of these connections can be considered as a feature extractor. Intermediate *I*-layer that consists of ON- and OFF-neurons is designed to work with grayscale images. The input of each *I*-layer neuron is connected to one neuron of the *S*-layer and the output is connected to the input of one neuron of the *A*-layer. All the *I*-layer neurons connected to one *A*-layer neuron form the group of this *A*-layer neuron. For example, in Fig. 3 the group of four *I*-layer neurons, two ON-neurons and two OFF-neurons, corresponds to one *A*-layer neuron. Reaction *R*-layer contains neurons that correspond to output classes of the LIRA classifier. Each neuron of the *A*-layer is connected to all the neurons of the *R*-layer. The weights of these connections are modified during the classifier training.

The fixed set of connections between the *S*-layer and the *A*-layer is created with the following procedure repeated for all the *A*-layer neurons: the window of height h and width w is randomly located in the *S*-layer; inputs of a group of *I*-layer ON- and OFF-neurons are connected to random neurons within the window $h \cdot w$ of the *S*-layer and outputs are connected to the *A*-layer neuron; the thresholds of ON- and OFF-neurons are selected randomly from the range $[0, b_{max}]$, where b_{max} is the maximal brightness of image pixels.

Each input image defines unique activations of the *A*-layer neurons. After the set of connections between the *S*-layer and the *A*-layer is created, the binary vector that represents outputs of associative neurons is calculated for each image of training and validation sets in accordance to following rules:

1. The output of the ON-neuron is equal to 1 if its input value is larger than its threshold and it is equal to 0 in the opposite case. The output of the OFF-neuron is equal to 1 if its input value is smaller than its threshold and it is equal to 0 in the opposite case.

2. The output of the *A*-layer neuron is equal to 1 if outputs of all the neurons of its *I*-layer group are equal to 1 and it is equal to 0 in the opposite case.

These binary vectors will be used during training and recognition procedures.

Training and recognition procedures of the LIRA neural classifier are similar to the ones of the perceptron. The training process is carried out iteratively. In each training cycle all the images of the training set are presented to the neural classifier. During the recognition process all the images of the validation set are presented to the neural classifier.

Image recognition performance of a LIRA neural classifier can be improved with implementation of distortions of input images during training and recognition [Kussul et al. 2006]. Different combinations of horizontal, vertical and bias translations of the spectrograms and scalograms were used in our experiments.

3.3. Experimental Results

Holdout cross-validation was used in our experiments, i.e. the validation set for each class was chosen randomly from the database and the rest of the database was used for training. In each experiment 50 runs of the holdout cross-validation were performed to obtain statistically reliable results. A new set of connections between the *S*-layer and the *A*-layer and a new division into the training and validation sets were created for each run.

The number of sounds in training and validation sets for each class equals ten, i.e. the database is divided in half.

The mean recognition rate was calculated from the mean number of errors for one run and the total number of sounds in the validation set. Comparison of recognition rates obtained with combination of LIRA with CWT and STFT for various numbers of associative neurons is presented in Table 1.

The following set of LIRA parameters was used during all the experiments: window $h \cdot w$ width $w = 10$, height $h = 10$; the number of training cycles is 30; the number of ON-neurons in the *I*-layer neuron group that corresponds to one *A*-layer neuron is 3, the number of OFF-neurons is 5; 8 distortions for training including ±1 pixel horizontal, vertical and bias translations and 4 distortions for recognition including ±1 pixel horizontal and vertical translations. Near optimal values of all the aforementioned parameters were determined empirically.

Table 1: Comparison of recognition rates for combination of LIRA with CWT and STFT

Number of associative neurons	Mean recognition rate (%)		*P*-value for paired *t*-test for mean recognition rate	95% lower bound for mean difference
	CWT	STFT		
1,000	85.3	81.75	0.02	0.72
2,000	96.5	94.25	0.002	1.039
4,000	99.6	98.1	< 0.001	0.926
8,000	100	99.85	0.042	0.0078

The paired *t*-test [Montgomery 2004] for mean recognition rate was used to evaluate the significance of the difference in recognition rates for CWT and STFT with null hypothesis of no difference in recognition rates and alternative of mean recognition rate for CWT being higher than the one for STFT. *P*-values and 95% lower bounds for mean

difference are presented in Table 1 indicating a statistically significant improvement in the recognition rate.

3.4. Discussion

The obtained results suggest the feasibility of the elimination of artifacts with an acoustical swallowing detection method as well as the superiority of the combination of LIRA with CWT over the combination of LIRA with STFT, though tests on a larger database would be needed for a conclusive proof. An important advantage of the proposed method is utilization of a double-redundant approach to identification of significant features. First, the time-frequency decomposition method provides a redundant description of a sound instance, therefore increasing the chances for random selection of a significant feature. Second, randomly assigned redundant connections between the sensor and associate layers ensure multiplicity of extracted random features. The proposed methodology eliminates the need for a separate feature selection and extraction algorithms and presents a novel deviation from the traditional approach of using small sets of empirically-selected statistics as features in sound recognition.

Higher accuracy achieved in classification of CWT data can be attributed to the tiling of the resolution. Time-frequency resolution of STFT is constant which results in the uniform tiling of the time-frequency plane with a rectangular cell of fixed dimensions. For CWT the time-frequency resolution varies according to the frequency of interest. CWT resolution is finer at higher frequencies at the cost of a larger frequency window while the area of each cell is constant. Hence, CWT can discern individual high

frequency features located close to each other in the signal, whereas STFT smears such high frequency features occurring within its fixed width time window [Addison 2002].

The method presented here achieves similar or higher accuracy compared to previously published methods. The advantage of this approach is that our method of feature extraction is automated. This has two main advantages. The first is that our method is not necessarily tailored to a specific collected dataset. That is, often in applications, features that are chosen manually from one dataset may achieve high performance for that dataset, but are not generalizable to the underlying application. The second is that manually chosen features may not achieve the best performance because potentially useful features for classification may have been overlooked. More research is needed to assess the generalizability and performance with this method of classifying swallowing sounds compared with other approaches.

The drawback of the proposed method is that it turned out to be very computationally intensive both in terms of data preprocessing and classification exceeding the capabilities of a typical low-power embedded processor and thus cannot be easily adapted for a wearable device for monitoring of the ingestive behavior. For example, computing CWT for a 20-minute long meal takes almost a day of processing on a computer equipped with an AMD Athlon 64 X2 4400+ Dual Core processor and 2.00 GB of RAM. Furthermore, the number of extracted random features which is equal to the total number of associative neurons is a crucial parameter of the system. This is reflected in the experimental results presented in Table 1. This number should be sufficiently large to create a detailed description of a sound instance providing a basis for further classification. At the same

time the large number of associative neurons results in a computational burden that may pose additional problems for our long term goal application. For example, in order to achieve an accuracy of more than 75% in automatic recognition on data from our final database (described in detail below) the number of associative neurons on the order of 500K was needed which makes recognition of swallowing instances in sound with the highest recognition rate in a real-time conditions problematic. Even with time-frequency decomposition spectra images resized to only 64x64 pixels using bicubic interpolation detection of swallowing instances on data representing approximately 20-minute long meals for 10 subjects took almost a week of processing on the aforementioned computer. Decreasing the size of the spectra resulted in average per-epoch accuracy of 93.06% obtained for individual models where the training dataset was selected in the following way: the total number of swallows per subject was rounded down to the closest number divisible by 10 and this divisible number of swallow spectra images was chosen randomly from the total number of swallow images with the same number of non-swallow images being chosen randomly from the total number of non-swallow images.

Given the slow performance and high computational burden of the combination of CWT with a LIRA neural classifier, a new light-weight sound recognition methodology based on a combination of Wavelet Packet Decomposition (WPD), mel-scale Fourier Spectrum (msFS) and Support Vector Machines (SVM) was proposed.

3.5. Conclusion

In this chapter a novel swallowing sound recognition technique based on the limited receptive area (LIRA) neural classifier and time-frequency decomposition was proposed.

The proposed technique works by applying a LIRA-based multipurpose image recognition system to the time-frequency decomposition spectrums of sound instances with extraction of a large number of random features. Features that do not provide useful information for separation of classes do not obtain significant weights during training. This approach eliminates the need for empirical feature selection and therefore simplifies the design of pattern recognition systems for non-stationary signals such as swallowing sounds.

The proposed methodology was tested with two different algorithms of time-frequency decomposition, short-time Fourier transform (STFT) and continuous wavelet transform (CWT), in recognition of four classes of sounds that correspond to swallowing sounds, talking, head movements and outlier sounds. Experimental results suggest the feasibility of the elimination of artifacts with an acoustical swallowing detection method, efficiency and reliability of the proposed method and the superiority of the combination of LIRA with CWT over the combination of LIRA with STFT. The drawback of the proposed method is its high computational burden.

The proposed multipurpose sound recognition technique may be employed in systems for automated swallowing assessment and has the potential for application to other sound recognition tasks.

4. Human Study and Data Collection Process

Related publications:

- Sazonov E, Schuckers S, Lopez-Meyer P, Makeyev O, Sazonova N, Melanson E, Neuman M (2008) "Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior" Physiological Measurement, 29:525-541.
- Sazonov E, Schuckers S, Lopez-Meyer P, Makeyev O, Sazonova N, Melanson E, Neuman M (2009a) "Reply to 'Comment on Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior'" Physiological Measurement, 30:L5-L7.

In this chapter, the human study that was performed to collect the data needed to validate the proposed swallowing and food intake detection methodologies is described. Major contributions made by the author of this dissertation to the data collection process included: initial testing of the set-up and several variations of microphones and strain sensors in different combinations, final selection of sensors for the study and creation of data collection protocol. The author also actively participated in conductance of the human study and data collection.

Multi-modal data collection system was designed for non-invasive monitoring of chewing and swallowing [Sazonov et al. 2008, Sazonov et al. 2009a]. Monitoring is based on detecting swallowing by a microphone located over the laryngopharynx or by a bone conduction microphone and detecting chewing through a below-the-ear strain sensor (Fig. 4). A strain sensor will detect specific motion of the lower jaw by capturing strains created by motion of the posterior border of the mandible's ramus relative to the temporal bone (Fig. 4).

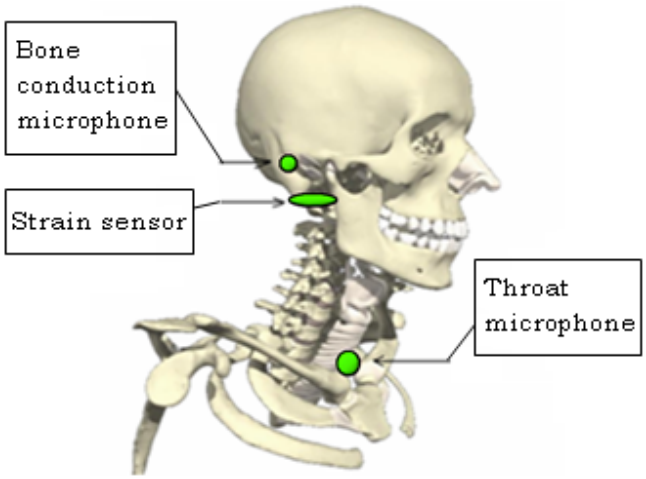


Figure 4: Suggested sensor locations.

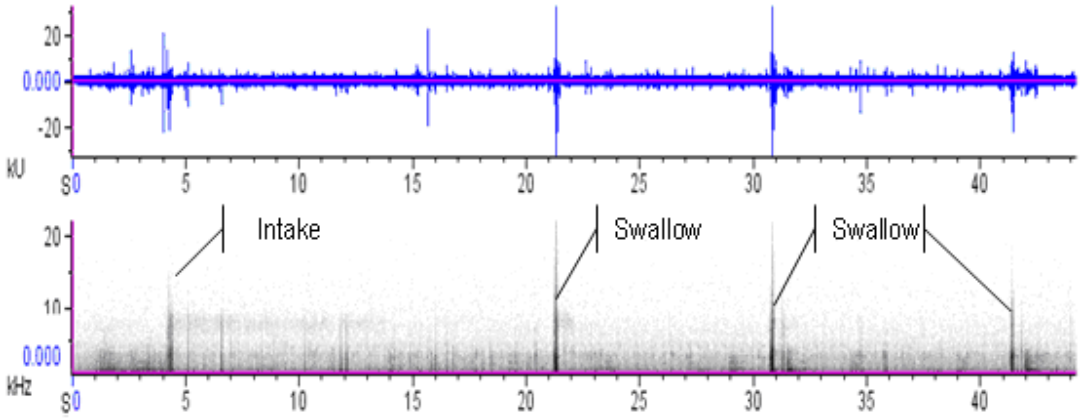
First, a system comprised of sensors, related hardware and software for multi-modal data capture was designed for data collection in a controlled environment. Second, a protocol was developed for manual scoring of bites, chews, and swallows for the collected data for use as a gold standard. The multi-modal data capture was tested by measuring chewing and swallowing in twenty one volunteers during periods of food intake and quiet sitting (no food intake). Video footage and sensor signals were manually scored by trained raters. An inter-rater reliability study for three raters conducted on the sample set of 5 subjects resulted in high average intra-class correlation coefficients of 0.996 for bites, 0.988 for chews, and 0.98 for swallows for the epoch duration of 120s. The collected sensor signals and the resulting manual scores were used as a gold standard in testing of the automatic swallowing detection methodology utilizing only the information from the wearable sensors and study of the relationship between swallowing/chewing and ingestive behavior.

4.1. Sensors

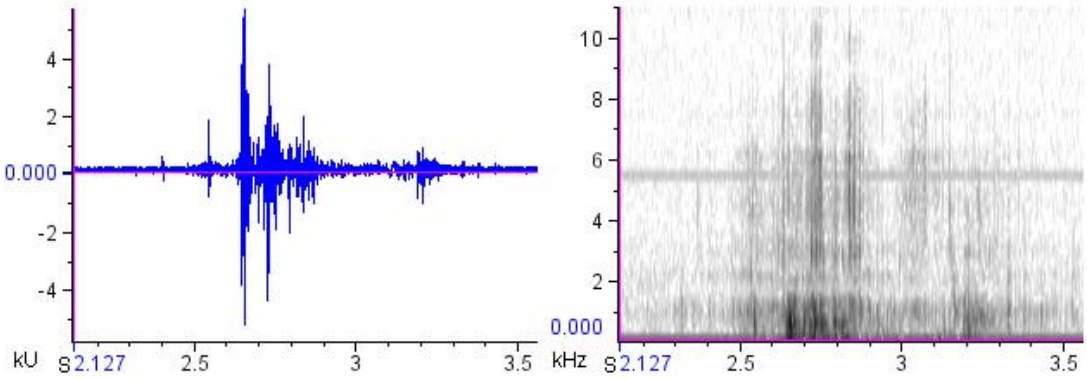
Four models of commercially available miniature microphones were tested as the sensing devices. The first model was a piezoelectric bone-conduction microphone EM-L (Temco Inc). This microphone can be modified to be placed on the mastoid bone behind the ear or used as an ear probe. The second model was a piezoelectric noise-canceling microphone model N4530 (Challenge Electronics). The third model was a modified throat microphone XTM70V (iXradio) usually used for hands-free radio communications. Throat microphones convert vibration signals from the surface of the skin rather than pick up waves of sound pressure, thus reducing the ambient noise. The fourth model was a miniature IASUS NT (IASUS Concepts Ltd) throat microphone. This microphone provides a dynamic range of 46 ± 3 dB with a frequency range of 20 Hz to 8000 Hz. Youmans (2003) reported the peak frequency of swallowing to be in the range of 1083.02 Hz to 3286.73 Hz, therefore this microphone is capable of acquiring swallowing sounds.

The microphone tests consisted of recording several consecutive swallows with subsequent evaluation of sound quality. Sound quality was evaluated both subjectively (i.e. by listening to the recording) and objectively (by visualizing in a sound editor and computing the signal-to-noise ratio). While all four microphones were able to detect swallowing sounds, the throat microphones showed a lower degree of sensitivity to ambient noise. Based on the results of testing, the IASUS microphone was selected for data collection because of its higher sensitivity to swallowing sounds and low noise (Fig. 5).

a)



b)



c)

Figure 5: a) IASUS NT throat microphone, b) time series and spectrogram of consuming 5 peanuts with 3 swallows, c) time series and spectrogram of a gulp of water of arbitrary size.

Several configurations of below-the-ear strain sensors for chewing detection were developed and tested. Evaluated types of sensors included foil strain gauges and a piezoelectric sensor.

Testing the strain sensors consisted of three distinct activities: orally counting to ten, drinking 50 ml of water and eating a cracker. Different sensors and configurations were evaluated with regard to sensitivity to the characteristic chewing motion and an ability to reject anterior-posterior and medial-lateral head tilts. While all sensor types are able to detect characteristic jaw motion due to chewing, the sensor configurations using foil strain gauges showed a higher degree of sensitivity to the subject's head motion which was especially evident for head tilting during drinking. Based on the results of the testing, a piezoelectric film sensor (MSI Inc) was selected to be used for the data collection. Attached by medical tape to the area immediately below the outer ear, this sensor is able to detect changes in the skin curvature created by the characteristic motion of the mandible relative to the temporal bone during chewing and bites.

The final set of sensors consisted of: (1) an IASUS throat microphone located over the laryngopharynx to detect swallowing, (2) a microphone directed outwards to detect ambient sounds, (3) a throat microphone located on the mastoid bone to detect swallowing, (4) a piezoelectric strain sensor attached by medical tape immediately below the outer ear to detect chewing, and (5) an in-ear microphone XEM98D (iXradio) to detect swallowing.

The block diagram of the system for multi-modal data collection is presented in Fig. 6.

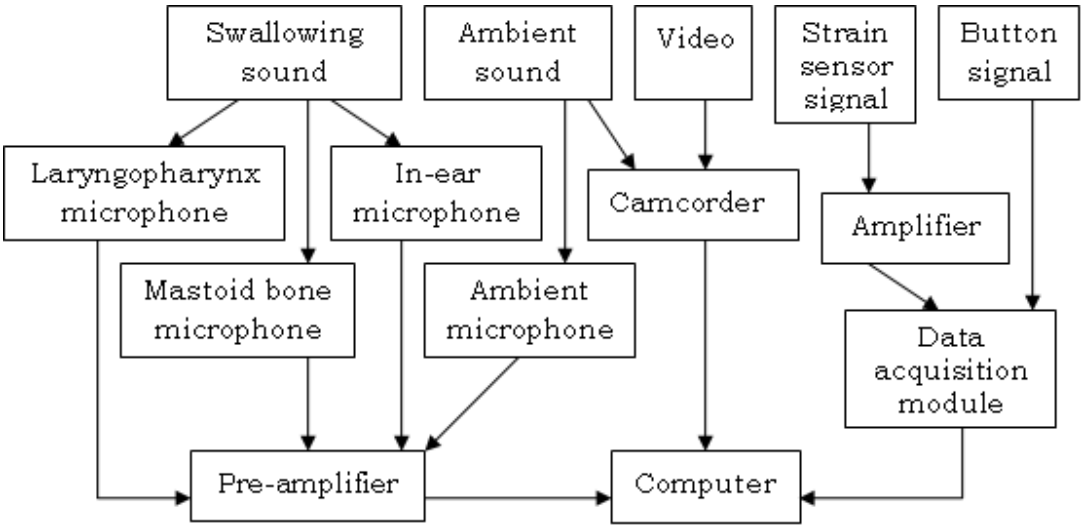


Figure 6: Block diagram of the multi-modal data collection system.

Microphone signals were amplified by a custom-built pre-amplifier with a variable gain in the range 20 dB to 60 dB. The gain of the amplifier was set experimentally for each sound channel to reliably capture the subtle sounds of swallowing without saturating the amplification circuits during normal speech and fixed for the whole data collection process. Amplified signals were recorded through a line-in input of a standard sound card at a sampling rate of 44100 Hz.

The signal from the piezoelectric strain gauge was buffered by a custom-designed amplifier with input impedance of approximately 100 MΩ. This buffered signal was acquired by a 16-bit data acquisition module USB-1608FS (www.measurementcomputing.com) at a sampling rate of 100 Hz.

A handheld push-button switch was connected to another input channel of the USB-1608FS. Subjects were asked to push the button to indicate swallowing instances which were recorded as a pulse of 5 V.

During each session of the data collection, subjects were videotaped in profile by a camcorder to capture subject activity and ambient sound independent of data acquisition by computer. Camcorder video was captured at 30 frames per second in an interlaced format. To simplify the scoring process, video was deinterlaced into a progressive 60 frames per second stream and the sound track was separated from video.

4.2. Software

Data acquisition software was developed in LabVIEW (National Instruments). The software allows simultaneous capture of 4-channel sound (from 2 sound cards) and up to 8 channels of sensor data (such as the strain sensor signal and the square wave from the button). All captured data are synchronized in time. Information about the data files and synchronization values was stored in a project file.

The scoring software (Fig. 7), also developed in LabVIEW, allows manual review and playback of the acquired data by a human rater and assignment of event marks to each instance of swallowing, each period of chewing with associated number of chews, and bites with associated mass of the bite.

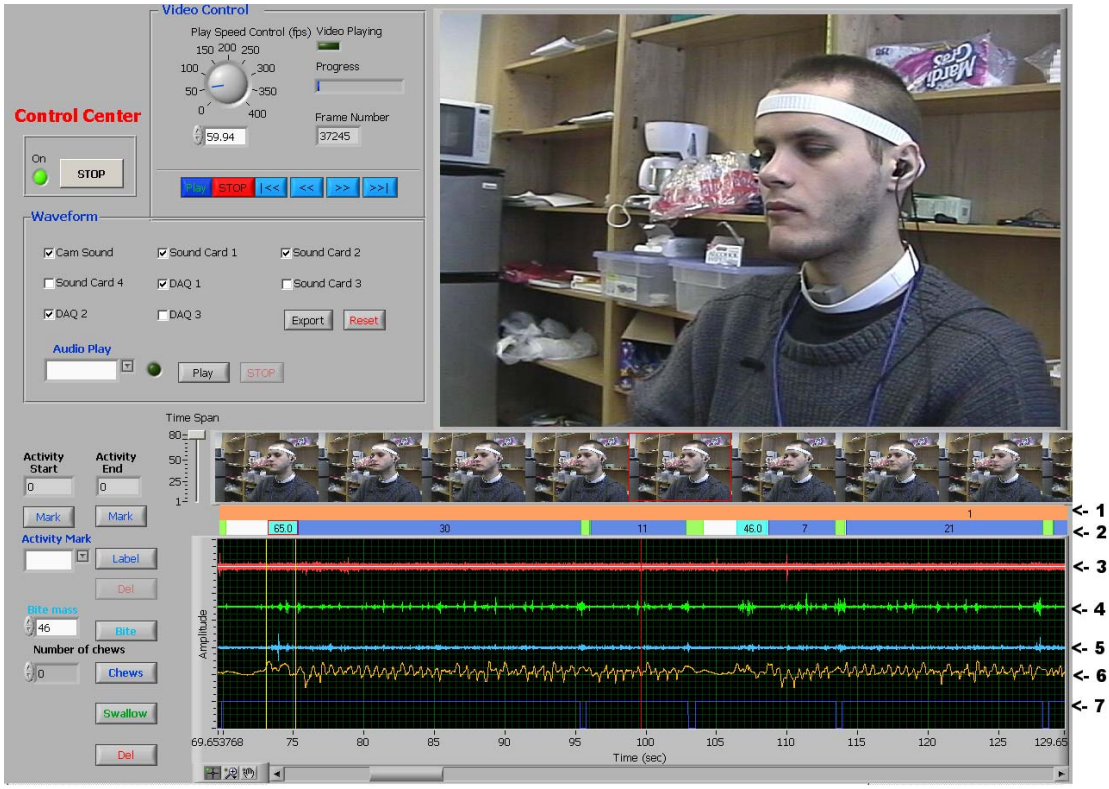


Figure 7: Scoring software graphical user interface: 1) activity mark, 2) bites/chews/swallows track, 3) ambient sound signal, 4) throat microphone signal, 5) bone conduction microphone signal, 6) strain sensor signal, 7) user button signal.

The scoring software also allows the user to zoom in and out in the data window, show all or selected data channels, and browse video frame-by-frame or any specified interval. The same software also allows assignment of labels for long-term activities performed by a subject. For example, periods of consumption of a specific type of food or a specific activity such as silent inactivity or talking can be indicated on the timeline. Manual scoring of the data utilizes all the data channels shown in Fig. 7, including the video and signals from the sound and strain sensors. A scorer following a predefined protocol identifies target segments of the time series, plays back the sensor data and narrows the boundaries of bites, chews and swallows.

4.3. Data collection protocol

Data collection was performed on a group of 21 generally healthy human subjects, 12 males and 9 females. In addition, since chewing and swallowing detection may be more difficult in obese individuals, thirty eight percent of subjects had a body mass index (BMI) greater than 30. The mean BMI of the subjects was of 28.98 with standard deviation of 6.42, subject's minimal BMI of 20.9 and maximum BMI of 42.1. Institutional Review Board approval was obtained for the study. Subjects read and signed the informed consent form. Data collection for each subject was performed during four visits. The subject's weight, waist and hip circumference (to identify android or gynoid type of obesity if present) were measured at each session. The subject's height was measured once during the first session. Subjects were encouraged to abstain from talking during the study unless they were asked to talk. All subjects had no dental problems that would interfere with normal food intake. As much as possible, an attempt was made to recruit a diverse population in terms of gender, ethnicity, age, and body size. However, due to the small sample size, at present, our sample may not be representative of the population.

Each session consisted of three parts: (1) a 20 min inactivity period (10 min of silent inactivity and 10 min of talking where the subject was asked to read aloud); (2) the meal period, consisting of unlimited time to eat a meal of a fixed size plus extra food items at the end of this period, if desired; (3) a second 20 min inactivity period (10 min of silent inactivity and 10 min of talking). A variety of magazines were provided to entertain the subject during the inactivity periods. Subjects were encouraged to read with a straight

neck, holding the magazine in front of the face to avoid obscuring the camcorder's perspective of the subject's neck.

Two fixed sizes of the meal (standard and large) were used with the large size being 50% bigger than the standard. The following food items were included in the meal: a slice of cheese pizza, a container of 1% fat yogurt, an apple, and a peanut butter sandwich. The foods were selected to represent different physical properties of the food such as crispiness, softness/hardness and tackiness. The variability in physical properties of food ensured that the proposed methodology was tested on a sample that is representative of the variability in the properties of everyday food. More analysis is needed to determine if sensors are capable of distinguishing between food properties. The provided drink was clear water. All food items were to be consumed unmixed and completely. The weight of the food item was measured after each bite on an electronic scale and recorded by the observer. Water was consumed separately from food.

During the first session, a standard size meal was served and no background noise was allowed during the meal period. During the second session, a standard size meal was served and background noise and talking to the subject were used during the meal period. Noise was introduced in the experiment to simulate realistic environments where people may be eating, and that can potentially impact results in future sound recognition experiments. To create background noise a combined recording of city noise, restaurant noise and segments of music recordings at a fixed volume level was used. To involve the subject in conversation, the operator asked the subject questions not relevant to the purpose of the research. During the third session, a large size meal was served and no

background noise was allowed during the meal period. During the fourth session, a large size meal was served and background noise and talking to the subject were used during the meal period.

The complete version of the data collection protocol is presented in Appendix A.

4.4. Conclusions

The human study described in this chapter was performed to collect the data needed to validate the proposed swallowing detection and food intake detection methodologies. A methodology of studying of ingestive behavior by non-invasive monitoring of swallowing (deglutition) and chewing (mastication) has been proposed based on data from sensors that may be implemented in a wearable monitoring device, thus enabling monitoring of ingestive behavior in free living individuals. The hardware/software system described in this chapter captures multi-modal sensor data which can be used for manual scoring of swallowing and food intake periods. These manual scores will further be used for assessment of swallowing and food intake detection accuracies.

5. Automatic Detection of Swallowing Events by Acoustical Means

Related publications:

- Sazonov E, Makeyev O, Schuckers S, Lopez-Meyer P, Melanson E, Neuman M (2010) "Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior" IEEE Transactions on Biomedical Engineering, 57:626–633.

Material of this chapter was published as a journal paper in [Sazonov et al. 2010] and is presented here in the corresponding form. The author of this dissertation was the primary student contributor to the work presented in this chapter.

5.1. Abstract

Our understanding of etiology of obesity and overweight is incomplete due to lack of objective and accurate methods for Monitoring of Ingestive Behavior (MIB) in the free living population. Our research has shown that frequency of swallowing may serve as a predictor for detecting food intake, differentiating liquids and solids, and estimating ingested mass. This paper proposes and compares two methods of acoustical swallowing detection from sounds contaminated by motion artifacts, speech and external noise. Methods based on mel-scale Fourier spectrum, wavelet packets, and support vector machines are studied considering the effects of epoch size, level of decomposition and lagging on classification accuracy. The methodology was tested on a large dataset (64.5 hours with a total of 9,966 swallows) collected from 20 human subjects with various degrees of adiposity. Average weighted epoch recognition accuracy for intra-visit individual models was 96.8% which resulted in 84.7% average weighted accuracy in detection of swallowing events. These results suggest high efficiency of the proposed

methodology in separation of swallowing sounds from artifacts that originate from respiration, intrinsic speech, head movements, food ingestion, and ambient noise. The recognition accuracy was not related to body mass index, suggesting that the methodology is suitable for obese individuals.

5.2. Introduction

The world is still losing in the battle with the obesity epidemic. According to WHO, in 2005 there were approximately 1.6 billion overweight and at least 400 million obese adults worldwide [WHO 2006]. Current trend is unsettling: 2015 projections predict 2.3 billion overweight and 700 million obese adults worldwide. Obesity is one of the risk factors for development of chronic diseases and presents a serious health problem. A recent study [Olshansky et al. 2005] suggested that effects of obesity on global health may be comparable to those of cancer. Though etiology of obesity is a topic of ongoing scientific debate, regulation of food intake may be the primary factor for maintaining a healthy weight [Flatt 1996] in the environment that provides abundance of inexpensive, highly palatable and energy dense foods, while requiring only minimal levels of physical activity [Hill et al. 2003].

While various methods have been developed for accurate and objective characterization of energy expenditure [Ainslie et al. 2003], at the present time, there is no accurate, inexpensive, non-intrusive way for objective Monitoring of Ingestive Behavior (MIB) in free living conditions. The most precise method of measuring energy intake is the Doubly-Labeled Water (DLW) technique which provides accurate estimates of caloric energy intake over long periods of time (10-14 days), if subjects remain weight stable.

However, the DLW technique cannot identify daily intake patterns. Dietary self-report methods like food frequency questionnaires [Weber et al. 2001], self-reported diet diaries [De Castro 1994], and multimedia diaries [Kaczkowski et al. 2000] have been shown to be inaccurate and underreport daily intake.

Our recent research [Sazonov et al. 2009b] has shown that frequency of swallowing events can serve as a predictor for accurate detection of food intake, differentiation between liquid and solid foods and estimation of ingested mass, with high frequency of swallowing being indicative of ingestion. Thus, an affordable wearable MIB device can potentially be created for objective capture and characterization of food intake. Such a device would capture both spontaneous and food intake swallows as they happen throughout the day without any conscience input from the user. A higher-level algorithm [Sazonov et al. 2009b] would detect and characterize food intake from the time series of swallows. Potentially, such a device can reduce underreporting because: 1) monitoring is objective and does not rely on self-report; 2) continuous capturing of spontaneous swallows indicates whether the sensor system is being worn or not, thus preventing or detecting intentional misreport.

While the weight gain is ultimately defined by the energy balance (energy intake minus energy expenditure) and the proposed MIB device by itself cannot capture the energy content of a meal, such a device can provide valuable information about ingestion that is not available at this time. Potentially, the device can help diagnose and treat dangerous behaviors leading to weight gain, such as unconscious snacking [Ward 1998], night eating [Stunkard 2002], and evening [Kant et al. 1995] or weekend overeating [Haines et

al. 2003]. The device may also find applications in diagnostics and treatment of disorders not directly related to obesity such as inadvertent weight loss (cachexia), anorexia and bulimia as well as dysphagia.

This paper presents a method for acoustical detection of swallowing events which is the first and fundamental step in implementation of the wearable MIB device. The swallowing detection does not need to differentiate between spontaneous and food intake swallows as the methods in [Sazonov et al. 2009b] rely only on frequency of swallowing events. Algorithms presented in [Sazonov et al. 2009b] can be applied as the second step of processing to detect and characterize food intake from the time series of swallows.

This paper demonstrates high accuracy of swallowing event detection by acoustical means on the largest dataset to date by the methodologies based on mel-scale Fourier Spectrum (msFS) and Wavelet Packet Decomposition (WPD) for time-frequency representation, and Support Vector Machines (SVM) for automatic recognition of characteristic sound of swallowing. It also contains assessment of the size of a near-optimal time decomposition window and effects of the decomposition level and epoch lagging on accuracy of swallowing detection suggesting that epoch duration used in earlier publications may not be optimal. Furthermore, assessment of recognition accuracy as a function of subject's Body Mass Index (BMI) shows that the proposed acoustical method is suitable for obese individuals. Finally, it is demonstrated that proposed methods have substantial tolerance to the sound artifacts resulting from food intake, intrinsic speech and background noise and thus may be suitable for free living conditions.

This paper is organized as follows: section 5.3 presents the background on assessment of swallowing sound signals and currently used automatic swallowing detection methods. Section 5.4 provides a brief description of the data collection process. Section 5.5 presents a detailed description of the proposed methodology. Experimental results are presented in section 5.6 followed by the Discussion and Conclusions.

5.3. Acoustical detection of swallowing events

At the present time videofluoroscopy and EMG are considered the gold standard in studies of deglutition. Videofluoroscopy depends on bulky and potentially unsafe equipment while EMG is too invasive due to frequently used subcutaneous placement of electrodes in the masseter, suprahyoid and infrahyoid muscles [Ertekin et al. 2002] to avoid interference from the muscles of the neck. Other reported sensors include a variety of strain devices [Ertekin et al. 2002, Stellar and Shrager 1985, Pehlivan et al. 1996]. However, most of the reported results indicate that detection of swallowing by a laryngeal strain sensor is not appropriate for obese subjects since under chin adipose deposits inhibit reliable detection of swallows. Use of accelerometer placed over the suprasternal notch of trachea as suggested by [Lazareck and Moussavi 2002, Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2006] may also be not appropriate for obese individuals for the same reasons. Detection of the characteristic swallowing sound created by the specific motion of laryngopharynx can be performed by a microphone which is significantly less invasive and more effective for obese individuals than the methods listed above.

Several methods have been proposed for assessment of swallowing sounds using signal processing and pattern recognition techniques. Papers [Lazareck and Moussavi 2002, Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2006, Aboofazeli and Moussavi 2008] presented methodologies for automatic decomposition of the tracheal sound signal into swallowing and respiratory segments in applications to dysphagia. The signal decomposition techniques utilized such features as autoregressive coefficients, root mean square values of the signal in time domain, average power of the signal within several frequency bands, waveform fractal dimension and Discrete Wavelet Transform on time windows (epochs) ranging in duration from 12.5 to 100 ms. Reported recognition rates were in the range from 78.54% [Lazareck and Moussavi 2002] to 93% [Aboofazeli and Moussavi 2006] although the sound recordings did not include any speech or noise.

Rejection of artifacts arising from ingestion, intrinsic speech and external noise is an issue that needs serious consideration. In the MIB applications, artifacts such as breathing, talking, throat cleaning, head movements, etc. may be confused with swallowing thus decreasing the accuracy of the recognition [Das et al. 2000]. The feasibility of sound artifact rejection was tested in [Makeyev et al. 2008b] where swallowing sound recognition was performed using the Limited Receptive Area neural classifier in combination with short-time Fourier transform and continuous wavelet transform. The methods in [Makeyev et al. 2008b] achieved 100% accuracy in classification of swallowing sounds on a limited dataset containing swallowing sounds, motion artifacts, talking and music, although practical applications to large datasets were limited by high computational burden of the method.

A recently reported method of automated swallowing detection that was tested in the presence of artifacts originating in talking, head movements, food ingestion, and respiration was presented in [Amft and Tröster 2008]. The data was collected from six healthy subjects using a sensor collar containing surface electromyography electrodes and a stethoscope electret microphone. A total of 7.93 hours of data with 1,265 swallows was acquired. Feature similarity search combined with an agreement of the detectors fusion method was used for classification. Four-fold cross validation was used with three folds used for training and one for validation. The average recognition rate of 70% was obtained for labeling epochs of 250ms as swallows/non-swallows but no accuracy in detection of swallowing events was reported.

In summary, acoustical detection of swallowing events, as presented herein, may present a non-invasive and convenient method suitable for use by obese individuals. However, the field of swallowing sound detection is relatively unexplored with a significant need to focus on realistic conditions with presence of various sound artifacts. Another key consideration is the choice of the epoch duration and lagging for signal analysis. Epoch sizes used in [Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2008, Amft and Tröster 2008] are substantially shorter (12.5-250 ms) than the average duration of a swallow (0.86 s) [Palmer et al. 1992] and thus may represent only a partial segment of a swallowing sound or require a large number of time lags. The goal of the methodology proposed in this paper is to consider acoustical swallow recognition as a method which may be appropriate for obese individuals; compare two popular signal time-frequency decompositions; investigate selection of key parameters of time-frequency transforms such as epoch duration and level of decomposition; and to test the proposed methods on a

challenging dataset that resembles free living conditions and includes artifacts of various origins.

5.4. Data collection

The data used in this paper were collected in human study reported in [Sazonov et al. 2008] where the details of the protocol, hardware, sensors and reliability of the manual scoring procedure are reported, but with no attempt to automatically recognize swallowing events. The following is a summary of the human study. The subject population included 20 volunteers, of which 7 had BMI greater than 30 (obese). Each subject participated in four visits, each of which consisted of a 20-minute resting period, followed by a meal, followed by another 20-minute resting period. Out of 80 collected visits, 10 were discarded due to data collection errors [Sazonov et al. 2008]. Selection and sequence of foods were fixed for each meal and represented different physical properties of the food such as crispiness, softness/hardness and tackiness, all of which may potentially impact both the artifacts arising from chewing sounds and the swallowing sound itself. To evaluate the impact of a meal-time conversation on the accuracy of swallowing detection, the subjects were involved in a dialogue with a member of the research team during the second and fourth visits and ate in silence during the first and third visits. Additionally, background noise (city noise, restaurant noise and music) were played during the second and fourth visits to simulate realistic environments where people may be eating. The subjects were monitored by a multi-modal sensor system which included an IASUS NT (IASUS Concepts Ltd) throat microphone located over laryngopharynx. The microphone provided a dynamic range of 46 ± 3 dB with a frequency range of 20 Hz to 8000 Hz. Amplified signals were recorded through a line-in

input of a standard sound card at a sampling rate of 44100 Hz. The recordings were manually scored to mark the boundaries of each swallow. The evaluation of inter-rater reliability reported in [Sazonov et al. 2008] showed high reliability of manual scores (0.98 average intra-class correlation) for detection of swallows.

5.5. Methodology

The methodologies proposed in this paper are based on two popular time-frequency decompositions: mel-scale Fourier Spectrum (msFS) and Wavelet Packet Decomposition (WPD) with classification performed by Support Vector Machines (SVM). Time-frequency decomposition and feature extraction based on WPD and msFS is widely used for processing of physiological signals, such as, for example, heart sounds [Turkoglu et al. 2003] and lung sounds [Liu et al. 2006, Cristianini and Shawe-Taylor 2000]. SVM is a supervised learning method that has a sound theoretical basis, is robust to overfitting (loss of generalization on noisy or incomplete data [Cristianini and Shawe-Taylor 2000]) and capable of producing very complex decision boundaries.

5.5.1. Feature Extraction by Wavelet Packet Decomposition

First, the sound stream was split into a series of overlapping epochs with fixed duration *D* and step *S*. A Hanning window was applied to each epoch. Second, a time-frequency decomposition of each epoch was obtained using Wavelet Packet Decomposition creating 2^N wavelet packets (where *N* is the level of decomposition) [Addison 2002]. A packet on the previous level is decomposed into two packets on the next level as:

$$w_{2n}(t) = \sqrt{2}\sum_k h_k w_n(2t-k) \qquad [5.1]$$

$$w_{2n+1}(t) = \sqrt{2}\sum_k g_k w_n(2t-k) \qquad [5.2]$$

where h_k is the low-pass Finite Impulse Response (FIR) filter and g_k is the high-pass FIR filter such as:

$$g_k = (-1)^k h_{1-k} \qquad [5.3]$$

The WPD was computed using Coiflet C4 wavelet. Advantages of the Coiflet wavelet include near linear phase, good amplitude response and fast computation [Fu et al. 2003]. WaveLab [Buckheit and Donoho 1995] package for Matlab was used to perform WPD. Third, each wavelet packet was converted into a scalar feature forming a feature vector f_i of length 2^N for each epoch. The chosen feature was the unbiased estimate of entropy [Moddemeijer 1989]. Fourth, to account for the time-varying structure of a swallow, a time-lagged feature vector was produced by merging feature vectors of the K adjacent epochs: $f_i^* = \{f_{i-K}, f_i, f_{i+K}\}$.

5.5.2. Feature Extraction by Mel-Scale Fourier Transform

First, segmentation of the sound signal into overlapping epochs was performed identically to the one used for WPD. Second, the Fourier amplitude spectrum *F(k)* of length L was computed for every epoch. Third, a mel-scale triangle filter bank $M_i(k)$ [Wu and Lin 2000] was used to compute 2^N point feature vector f_i (where N is an equivalent to WPD's level of decomposition) defined as:

$$f_i = \log\left(\sum_{k=0}^{L/2} F(k)M_i(k)\right), \quad i = 0,\ldots, N-1 \qquad [5.4]$$

Finally, the time-lagged vector f_t' was obtained in the same way as for WPD. Fig. 8 shows a segment of the sound recording containing a swallow and its respective representation obtained by WPD and msFS processing with decomposition level N = 8, epoch duration D = 1.5 s and step S = 0.2 s.

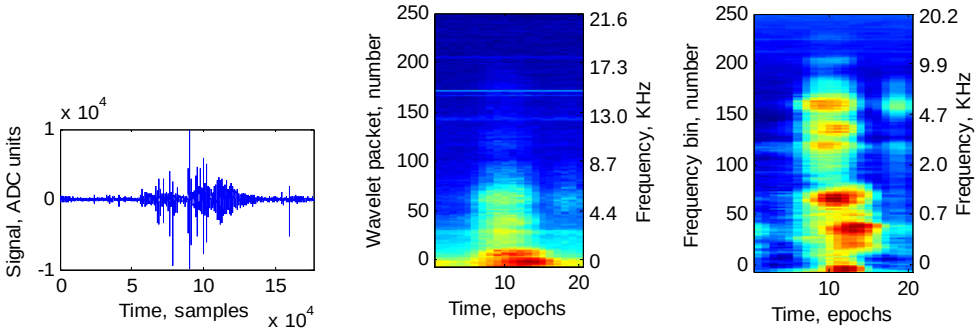


Figure 8: Feature extraction: a) A 4.0s fragment of a sound recording including a swallow, b) features extracted by WPD processing, c) features extracted by msFS processing. Frequencies are shown for the center of the extracted band.

5.5.3. Support Vector Machines

The time-lagged feature vectors f_t' obtained either through WPD or msFS processing were used as inputs for training and validation of an SVM classifier [Cristianini and Shawe-Taylor 2000]. The choice of the SVM as a classifier was defined by sound theoretical foundation and robust performance of SVM classifiers. A comparison of SVM performance to performance of 16 classification and 9 regression methods on 21 data sets for classification and 12 data sets for regression [Meyer et al. 2003] ranked SVM as one the best techniques on most data sets, especially for classification. LibSVM package for Matlab [Chang and Lin 2001] was used for training the SVM classifier using the Gaussian radial basis kernel function. Optimal parameters of the SVM classifier were found by a grid search procedure.

5.5.4. Selection of Optimal Epoch Duration and Decomposition Level

Optimal epoch duration D, epoch step size S, decomposition level N and number of lags K were determined in a grid search procedure. The epoch duration and step size (D/S) were taken from a set {3.0/0.4, 1.5/0.2, 0.75/0.1, 0.375/0.05} seconds which represents progressively finer time resolutions. Decomposition level both for WPD and msFS was taken as $N \in \{5,6,7,8,9\}$ thus producing from 32 to 512 features for each epoch. The number of lags K was either 0 or 1 since a higher number of lags produced long feature vectors which substantially slowed the classifier. Since a grid search procedure is time consuming, it was performed on randomly selected two visits that included noise and talking during the meal and thus presented a harder classification case. The grid search procedure repeatedly trained classifiers defined by various combinations of D/S, N and K. The validation accuracy was used to evaluate the goodness of parameters. Training and validation were performed with 34% of the data (one fold) used for training and 66% (two folds) used for validation. The accuracy of swallowing detection was estimated as described further.

5.5.5. Training and Validation

The pairs of feature vectors and class labels to be used as inputs for the SVM classifier were obtained in the following way: if any part of the epoch belonged to a swallow marked in the manual score the epoch label was set as ‘1’ (swallow epoch), otherwise it was set as ‘-1’ (non-swallow epoch). Individual intra-visit models were built for 70 visits of 20 subjects. The training and validation sets were formed by taking into account the highly non-homogeneous structure of each visit. For example, a period of quiet resting with no talking and no food intake will not have enough variability in the data to train a

classifier that would work reliably if talking or food intake is introduced. Since talking, food intake and external noise are introduced at various times in each visit, a longitudinal segmentation was used. Each visit was divided into 55 segments of equal duration, each segment 1 minute in duration on average. Three-fold cross-validation was performed with two folds used for training and one fold used for validation.

5.5.6. Accuracy of Detecting Swallowing Instances

Predicted class labels represent accuracy of the classifier on epoch level and do not correspond well to the accuracy of detection of swallowing events. Transition from the epochs to swallowing events was done by identifying all situations where either Manual Score (MS) or Automatic Score (AS) indicated presence of a swallow and calculating the numbers of true positives, false positives and false negatives in terms of swallowing events. A true positive (T_+) was counted if both MS and AS contained continuous sequences of epochs marked as swallows intersecting at one or more epochs or on the sequence boundary (Fig. 9, *a*). A false positive (F_+) was identified if the AS marked a swallow which was not present in the MS (Fig. 9, *b*). A true negative (T_-) was counted if both MS and AS contained continuous sequences of epochs marked as non-swallows intersecting at one or more epochs (Fig 9, *c*). A false negative (F_-) was counted if the MS marked a swallow which was not present in the AS (Fig 9, *d*). The accuracy of swallowing events detection was then estimated using weighted accuracy, sensitivity and specificity:

$$A = \frac{T_+ + T_-}{T_+ + T_- + F_+ + F_-} \qquad [5.5]$$

$$Sensitivity = \frac{T_+}{T_+ + F_-} \qquad [5.6]$$

$$Specificity = \frac{T_-}{T_- + F_+} \qquad [5.7]$$

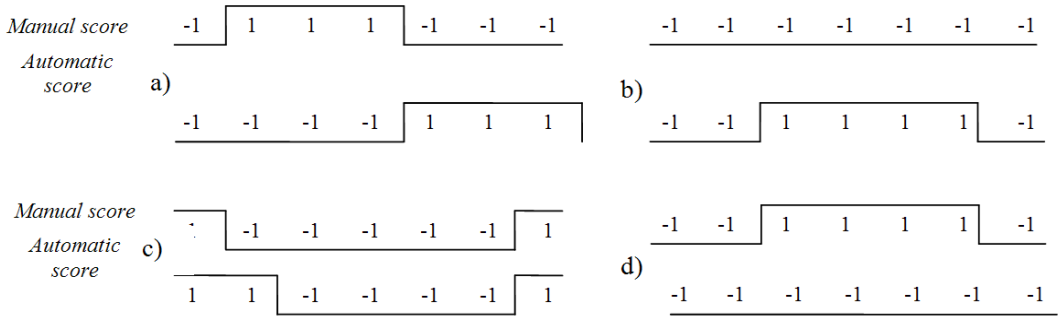


Figure 9: Examples of: a) true positive, b) false positive, c) true negative, d) false negative. Each number represents a class label for an epoch ('-1' – non-swallow epoch, '1' – swallow epoch).

5.6. Results

The graphs obtained by the grid search of optimal epoch duration, decomposition level and number of lags on a subset from 2 visits are shown in Fig. 10 which suggests the best parameters for WPD processing: 9th level of decomposition on 1.5 s epochs. For msFS processing the best parameters are at 7th level of decomposition on 1.5 s epochs. These parameters with and without lagging were used to process throat microphone signal collected in 70 visits. SVM training was performed with misclassification penalty C = 10 and Gaussian kernel width parameter $\gamma = 0.05$ obtained by a grid search. Results obtained in per-epoch recognition and detection of swallowing events are presented in Table 2.

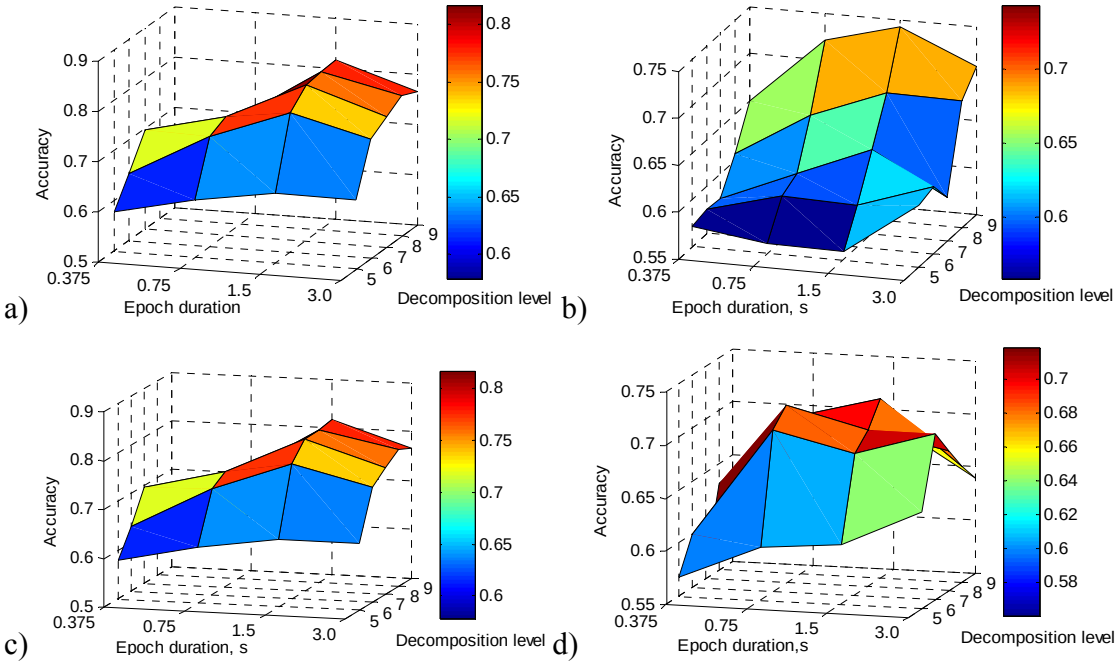


Figure 10: Accuracy of swallowing sound recognition as a function of epoch duration and decomposition level: a) msFS with no lags, b) WPD with no lags, c) msFS with K=1, (3 lags), d) WPD with 3 lags.

Table 2: Accuracy obtained in swallowing detection for three-fold cross-validation.

Feature	WPD-9	WPD-9	msFS-7	msFS-7
Number of lags	1	3	1	3
Average per-epoch accuracy (%)	95.9	96.4	96	96.8
Average per-swallow accuracy (%)	79.4	79.4	79	84.7

The best average weighted accuracy in terms of epochs and swallows was produced by msFS-7 with 3 lags and found to be 96.8±1.4% for epochs and 84.7±6.9% for swallows. The distribution of average weighted accuracy in classification of epochs and swallowing events versus the subject's BMI and corresponding linear fit of the data are presented in Fig. 11. To assess the impact of sound artifacts on accuracy of identifying swallowing events the average weighted swallowing accuracy was also computed individually for the four non-overlapping parts of the validation set corresponding to the following categories: periods of no food intake and no talking

(88.0%), periods of no food intake with talking (86.4%), periods of food intake and no talking and background noise (86.2%), and periods of food intake with talking and background noise (82.9%).

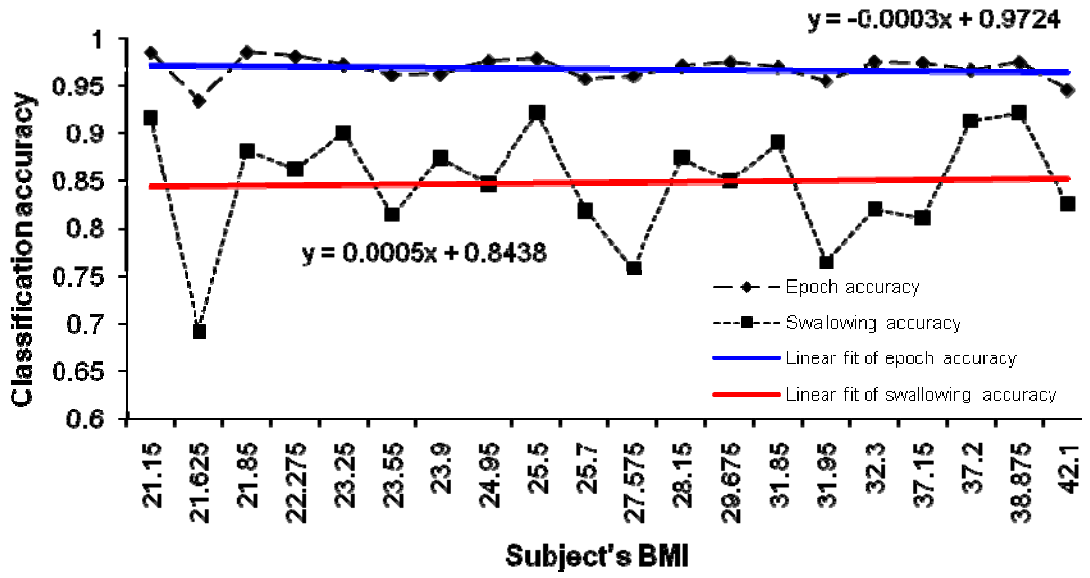


Figure 11: Distribution of average weighted accuracy in classification of epochs and swallowing events versus subject's BMI and corresponding linear fit of the data.

5.7. Discussion

One of the goals of this work was to determine the optimal duration of an epoch since durations reported in existing literature [Aboofazeli and Moussavi 2004, Aboofazeli and Moussavi 2008, Amft and Tröster 2008] varied over a wide range of 12.5-250 ms. As Fig. 10 shows that the epoch duration of 1.5s clearly demonstrates the highest recognition accuracy both for msFS and WPD with or without lagging. This corresponds well with the mean duration of swallow which in our study was found to be 1.15 s with a standard deviation of 0.29 s (based on analysis of 10,686 swallows), comparable to previously reported duration of 0.86s [Palmer et al. 1992]. Thus, the epoch duration of 1.5 s is sufficient to completely include an average swallow. We believe that such choice of the

epoch duration is one of the reasons that our recognition rate is substantially higher than per epoch accuracy of 70% reported in [Amft and Tröster 2008] where the authors used a 0.25 s epoch which cannot cover a complete swallow. Although our study excluded dysphagic subjects, we may anticipate that recognition of longer-than-normal dysphagic swallows [Vaiman and Nahlieli 2009] may benefit from a longer epoch.

Fig. 10 also demonstrates accuracy growth with increase in the level of decomposition. The most pronounced increase for msFS is observed up until the 7^{th} level of decomposition. As Table 2 shows a lagged version produces higher overall accuracy due to better preservation of accuracy during transitioning from epochs to swallows on some of the visits. Lagging takes the feature evolution over time into account and thus produces more accurate results. The non-lagged version of WPD processing clearly peaks at the 9^{th} level of decomposition and trends toward further growth. Unfortunately, higher levels of decomposition result in unacceptably long processing times both for feature extraction and classification. The lagged version of WPD behaves somewhat erratically (which may be attributed to a limited dataset used in the grid search procedure) but clearly peaks at 8^{th} level of decomposition, confirming that 8^{th}-9^{th} levels are probably near the optimum for WPD. Fig. 10 and Table 2 also demonstrate that msFS time-frequency decomposition clearly outperforms WPD resulting in higher recognition accuracy. A possible explanation is non-linear scaling of the frequencies by msFS which allows for a better representation of the lower frequencies which contain most of the energy of a swallowing sound.

One of major advantages of the current study is that it was designed to be close to real life conditions and include sound artifacts originating from chewing of food of different textures, talking, head movements, occasional intrinsic sounds (for example, coughing), and background noise of various origins. Thus, the classifier had to deal with a significantly more complex problem than previous studies while achieving a comparable (vs. 93% in [Aboofazeli and Moussavi 2006]) or better performance (vs. 79% in [Aboofazeli and Moussavi 2004]). The closest study that allows direct comparison is [Amft and Tröster 2008] which achieved the epoch-accurate average recognition rate of 70% but did not report on accuracy of detecting swallowing events. For comparison, the methodology proposed in this paper yielded the average weighted epoch accuracy of 96.8% that relates to 84.7% average weighted accuracy in detection of swallowing events. Furthermore, our study utilized a wider variety of solid foods (cheese pizza, an apple, and a peanut butter sandwich) with varying physical properties that directly impact the sounds of mastication [De Belie et al. 2003] and subsequently influence swallowing recognition. Another advantage is unrestricted consumption of liquids which were limited in [Amft and Tröster 2008] to 5ml and 15ml of volume at a time. Liquid consumption is characterized by a very high swallowing frequency [Sazonov et al. 2009b] in which identification of individual swallows is difficult due to the fact that consecutive swallows may be recognized as one. The results show that artifact sounds negatively impact the recognition accuracy but not to a degree that would render the method unusable. As expected, the highest recognition accuracy is observed for quiet periods of no food intake (88.0%) and the lowest recognition accuracy is observed for periods of food intake

combined with talking and background noise (82.9%). Thus, application of noise cancellation techniques may further improve on the classification accuracy.

As Fig. 11 suggests the accuracy of detecting swallowing events is very likely not to be dependent on the subject's BMI. While more data is needed to obtain higher statistical significance, this may be an important advantage of the acoustical approach of detecting swallowing events. The highest BMI of a volunteer in the study was 42.1 which is considered severe (morbid) obesity. Even for this volunteer the swallowing identification accuracy was greater than 80%. Thus, these results suggest that the proposed methodology could be used for monitoring of food intake in obese individuals.

The reported experimental results were obtained on the dataset containing 64.5 hours of data with 9,966 swallows collected from 20 subjects with the experimental conditions resembling those of food consumption in free living. To our knowledge this is largest dataset collected to date. In addition, the manual score of swallows used for training of the classifiers has known reliability metrics [Sazonov et al. 2008]. Overall, the proposed methodology showed good performance in testing on a more complicated dataset than any of the previous studies. The next step in the development of the acoustical method of the detection of swallowing is development of inter-visit individual and group models that could be practically applied for automatic scoring of the swallowing sound recordings. The desired accuracy of the identification of swallowing is another question that needs further investigation. However, the methods for detection of food intake and prediction of ingested mass [Sazonov et al. 2009b] should offers some tolerance to the

errors in detection of swallowing instances since they rely on multiple swallows and relatively long time windows (up to 2 minutes).

The results of this study also have important implications for the original intent to use automatic recognition of swallowing sounds in a wearable device for monitoring of ingestion. The time sequence of swallows detected by the proposed method can be further processed by algorithms in [Sazonov et al. 2009b] to detect and characterize food intake to achieve real-time monitoring of ingestion. With the rapid progression of computing power available in modern ubiquitous platforms (cell phones, PDA) the proposed MIB methodology can be implemented as a wearable device allowing for real-time biofeedback to individuals. Such a wearable device may potentially find numerous applications in research, clinical nutrition and self-monitoring of food intake by general population.

5.8. Conclusion

In this paper we describe two automatic acoustical swallowing detection methods for use in MIB applications. The methods were based on combination of mel-scale Fourier Spectrum (msFS) or Wavelet Packet Decomposition (WPD) and Support Vector Machines. The proposed methodology was tested on the data collected from 20 human subjects with 35% of the subjects being obese with Body Mass Index (BMI) of at least 30 and the average BMI of 28.53 using a multi-modal data collection system designed for non-invasive monitoring of chewing and swallowing. The total duration of data used for training and validation was 64.5 hours including 9,966 swallows which makes it the largest dataset to date. Average weighted epoch classification accuracy of 96.8% resulted

in 84.7% average weighted accuracy in detection of swallowing events. Optimal duration of a sound time slice was found to be 1.5s which corresponds well to statistics of swallowing duration. The msFS decomposition with 3 lags clearly outperformed WPD in recognition accuracy. A study of impact of food intake, talking and background noise on accuracy of swallowing detection suggests robustness of the proposed methodology to such events as well as its ability to accurately separate swallowing sounds from sound artifacts that originate in respiration, talking, head movements, food ingestion, and ambient noise. The method was also demonstrated to work equally well for both obese and non-obese subjects. The described methodology and sensors may be implemented in a wearable monitoring device, thus enabling MIB applications in free living individuals.

6. Automatic Food Intake Detection Based on Swallowing Sounds

Related publications:

- Makeyev O, Lopez-Meyer P, Schuckers S and Sazonov E (2010) "Automatic food intake detection based on swallowing sounds" Physiological Measurement (In review).

Material of this chapter was prepared for publication and submitted as a journal paper [Makeyev et al. 2010] and is presented here in the corresponding form. The author of this dissertation was the primary contributor to the work presented in this chapter.

6.1. Abstract

This paper presents a novel fully automatic food intake detection methodology, an important step toward objective monitoring of ingestive behavior. The aim of such monitoring is to improve our understanding of eating behaviors associated with obesity and other eating disorders. The proposed methodology consists of two stages. First, acoustic detection of swallowing instances based on mel-scale Fourier spectrum features and classification using support vector machines is performed. Principal component analysis and a smoothing algorithm are used to improve swallowing detection accuracy. Second, the frequency of swallowing is used as a predictor for detection of food intake episodes. The proposed methodology was tested on data collected from 12 subjects with various degrees of adiposity. Average accuracies of >80% and >70% were obtained for intra-subject and inter-subject models correspondingly with a fine time resolution of 30s. Results obtained on 44.1 hours with a total of 7305 swallows show that detection accuracies are comparable for obese and lean subjects. They also suggest feasibility of

food intake detection based on swallowing sounds and potential of the proposed methodology for automatic monitoring of ingestive behavior. Based on a wearable non-invasive acoustic sensor the proposed methodology can potentially be used in free-living conditions.

6.2. Introduction

This paper extends our work on development of automatic and objective approach to monitoring of ingestive behavior (MIB) in free-living conditions based on data from wearable non-invasive sensors [Sazonov et al. 2008, Sazonov et al. 2009a, Sazonov et al. 2009b, Sazonov et al. 2010]. Such an approach can be helpful in characterization of ingestive behaviors associated with a variety of eating disorders and for development of clinical interventions. The World Health Organization predicts 2.3 billion overweight and 700 million obese adults worldwide by 2015 [WHO 2006] and MIB could potentially be used in active weight control programs providing the objective feedback needed for diet management [Sazonov et al. 2009b, Amft and Tröster 2009]. Objectivity of such feedback is crucial as unhealthful and extreme weight-control were shown to predict outcomes related to obesity and eating disorders [Neumark-Sztainer et al. 2006].

Most of currently used self-reporting techniques demonstrate widespread bias to the underestimation of food intake [Livingstone and Black 2003] Because of bias and imprecision, self-reported food intake should be interpreted with caution unless independent methods of assessing its validity are included in the experimental design [Schoeller 1995]. Replacing paper-based reports with manually operated electronic devices to simplify tedious and error-prone logging did not improve validity of the

reporting [Yon et al. 2006]. A potential solution is to replace or augment manual self-reporting where individuals have to record their own eating behavior with automatic and objective sensor based monitoring where eating behavior is estimated without individual's active participation. This could significantly improve the accuracy reducing intake underreporting and relieving the individual from the recording burden while non-invasiveness and wearability of MIB sensors ensure their suitability for long-term monitoring in free-living conditions.

MIB-based characterization of food intake behavior includes several dimensions including: detection of periods of food intake, differentiation of solid foods from liquids, recognition of food type, prediction of the mass of ingested food and evaluation of caloric intake [Sazonov et al. 2009b]. In this paper we concentrate on development of objective and automatic approach to detect periods of food intake as a step towards our long-term objective to create an automatic, non-invasive and wearable MIB device suitable for use in free-living conditions. For this purpose we proposed a sensor system for non-invasive monitoring of chewing and swallowing, validated the reliability of the produced manual scores [Sazonov et al. 2008, Sazonov et al. 2009a], established a methodology of automatic detection of swallowing instances by acoustical means [Sazonov et al. 2010], and developed a methodology for detection and characterization of food intake based on manual scores of chewing and swallowing [Sazonov et al. 2009b].

This paper makes the next fundamental step toward objective MIB by integrating and validating a fully automatic food intake detection methodology based on acoustical detection of swallowing. The proposed methodology consists of two stages: first, acoustic

detection of swallowing instances based on mel-scale Fourier spectrum features and classification using support vector machines is performed. Principal component analysis and smoothing algorithm are used to improve swallowing detection accuracy. Second, frequency of swallowing is used as a predictor for detection of food intake episodes. Scheme of the proposed methodology is presented in Fig. 12.

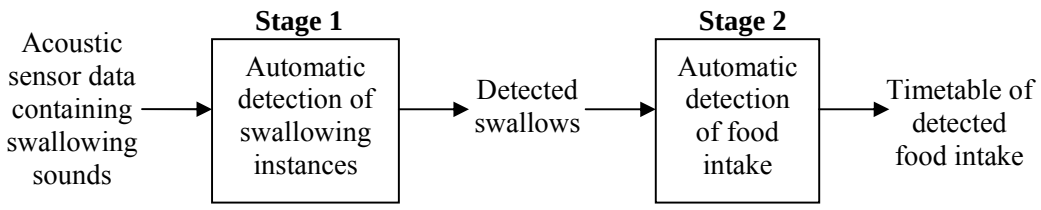


Figure 12: Scheme of the two-stage automatic food intake detection.

The proposed methodology was tested on large dataset (44.1 hours with a total of 7305 swallows) collected from 12 subjects with various degrees of adiposity.

Detection of periods of food intake was performed for both intra- and inter-subject models that can be directly implemented in a wearable MIB device. Average accuracies of >80% and >70% were obtained for intra-subject and inter-subject models correspondingly with a fine time resolution of 30s. Detection accuracies are comparable for obese and lean subjects and suggest feasibility of food intake detection based on swallowing sounds. To the best of our knowledge this is the first attempt of fully automatic detection of food intake based on swallowing data from wearable non-invasive sensors. Detailed review of previous attempts to detect food intake such as gesture based approach proposed in [Junker et al. 2008] and a chewing based approach proposed in [Nishimura and Kuroda 2008] is presented in section 6.3 as well as a review of related

work on automatic food intake detection based on data from non-invasive wearable swallowing sensors. The rest of this paper is organized as follows: description of the proposed methodology and the dataset used for its validation are presented in section 6.4. Experimental results for intra- and inter-subject food intake detection models are presented in section 6.5 followed by Discussion and Conclusions.

6.3. Related Work

Three categories of non-invasive wearable sensors have been proposed as basis for creation of automatic food intake detection methodology: intake gesture sensors [Junker et al. 2008], chewing sound sensors [Nishimura and Kuroda 2008] and swallowing sensors [Sazonov et al. 2009b, Sazonov et al. 2010]. Activities that correspond to these sensor categories represent a temporal description of food consumption and thus can be used to identify periods of food intake. A detailed review of these sensor categories in terms of their suitability for detection of periods of food intake is presented below.

Intake gestures are intentional upper body (arms and trunk) movements related to food intake. Compared to a simpler task of recognition of isolated movements, the task of gesture detection is more difficult because relevant gestures occur sporadically in a continuous stream of data while being embedded into other, partly arbitrary movements that are difficult to model due to their complexity and unpredictability [Junker et al. 2008]. Other sources of challenges include: co-articulation, where consecutive relevant gestures influence each other, and intra- and inter-person variability, e.g. in gesture duration. All this applies to the task of intake gestures detection solving which would reveal information about the timing of nutrition events providing an estimate of periods

of food intake. At this moment, the only approach to this task was proposed in [Junker et al. 2008] based on data from five inertial sensors attached to wrists (2), upper arms (2) and upper torso. This two-stage approach combines natural partitioning based pre-selection with Hidden Markov Models based classification. It was tested on frequently used human feeding movements of both arms and the trunk including: using fork and knife for intake of Lasagne, using spoon for intake of cereals or soup, drinking from a cup, and intake of bread or chocolate bar using one hand only. The data was collected from four subjects; two sessions were performed for each subject on different days. Total of 4.7 hours long of data was collected with 34.7% of the data containing intake gestures. Average recall of 0.78 and average precision of 0.77 were obtained for subject-specific or intra-subject prediction models. No results for non-personalized inter-subject model were reported. Even though obtained accuracy suggests the potential of the approach for food intake detection it has significant limitations. First, not all the food items require intake gestures, e.g. a high-caloric milkshake can be consumed using a straw. Second, arm movements to the head that are not related to food intake, e.g. brushing teeth, smoking, etc, can result being a significant source of misclassifications. Finally, the authors report high intra-subject variability of intake gestures caused by: differences in size and consistency of food pieces, temporal aspects, such as changes in food temperature and natural satiety subjects were developing during the intake sessions. Even though inter-subject variability of intake gestures wasn't evaluated it may be significant taking into account differences in human eating behaviors, e.g. eating with chopsticks versus cutlery. Therefore, a sensing solution for activity with smaller intra- and inter-subject variability

may be a better choice for current task, e.g. chewing and swallowing that seem to be less related to personal eating habits compared to intake gestures.

The task of food intake detection based on chewing is virtually identical to the task of detection of chewing instances as chewing sequence which usually starts right after the food piece is transferred to the mouth. The crushing of food and mixing it with saliva to form a bolus for swallowing is performed with cyclic opening and closing of the jaw and arbitrary tongue movements. An approach for automatic detection of chewing instances in a continuous data stream produced by a wearable non-invasive sensor was proposed in [Nishimura and Kuroda 2008]. In this approach wireless in-ear microphone is used to capture sound emissions generated by chewing and transmitted by bone conduction to the ear canal. Two-staged chewing detection algorithm first detects chew-like signals by applying number of zero-crossings threshold to log energy regression coefficients. Then chewing sound verification is performed based on similarity of signals detected at the first stage to the chewing sound models derived from the training data. High average chewing detection accuracy of 98.7% is reported for five food categories including chips, salad, rice, wafers and banana. However, limited data about the method validation are provided including the average number of test chews (516) and the number of training chews (100) per food category. It is not clear how many human subjects participated in the study, whether intra- or inter-subject model results are reported, how the data was divided into training and validation sets, what kind of validation technique was used, etc. Furthermore, usability of food intake detection based on chewing sensor is limited to solid foods since there is little to no chewing present during consumption of liquid and certain semisolid (yogurt, pudding, etc) food items. Such a limitation makes chewing

sensor a feasible candidate for sensor fusion rather than for an independent food intake detection sensor. Finally, absence of spontaneous chewing throughout the day as compared to, for example, swallowing gives no indication whether the MIB device based on the chewing sensor is being worn or not making the device vulnerable to intentional misreport of food intake.

Unlike chewing, swallowing occurs sporadically and unconsciously throughout the day so automatic detection of swallowing instances is only the first step towards the swallowing based food intake detection. Detected swallows further need to be classified as either spontaneous or food intake swallows. Our work on creation of automatic food intake detection methodology based on swallowing started with development of non-invasive multi-modal monitoring system including a wearable acoustic swallowing sensor – a throat microphone located over laryngopharynx [Sazonov et al. 2008, Sazonov et al. 2009a]. This monitoring system comprised of hardware, software and protocol for manual scoring of the collected data was used in a human study measuring chewing and swallowing in 21 subjects during food intake and resting periods. Reliability of the manual scoring process was validated using inter-rater reliability study conducted on sample set of five subjects for three raters. For swallowing scores high intra-class correlation coefficient of 0.98 was obtained suggesting that manual scores are reliable to be used as gold standard for validation of automatic swallowing and food intake detection algorithms on this large dataset of over 65 hours of data with over 10K swallows. Next, in [Sazonov et al. 2009b] we proposed and validated methodology for detection and characterization of food intake based on manual swallowing scores. In particular, we showed that instantaneous swallowing frequency defined by inverted difference in time

between consecutive swallows can serve as a predictor for accurate detection of food intake with average accuracy of 87% achieved for 30s epochs. We also proposed and validated methodology for automatic detection of swallowing instances by acoustical means [Sazonov et al. 2010] yielding 84.7% average accuracy in detection of swallowing events for intra-visit individual models. These results suggest the potential of using the swallowing frequency for automatic detection of food intake. However, two crucial questions remained unanswered. First, methodology of automatic swallowing detection was only validated on intra-visit individual models, i.e. training and validation of the algorithm were performed on different segments of the same recording rather than on separate recordings from the same or different human subjects, that can not be implemented in a MIB device as constant real-time re-training of the detection model in free-living conditions is not feasible due to the absence of gold standard manual score. Therefore, further validation is needed for intra- and inter-subject models which can be preprogrammed and implemented in such a device directly. Second, food intake detection methodology was only validated on manual scores. Validation on automatically produced swallowing scores is needed to evaluate how sensitive the food intake detection algorithm is to errors in swallowing detection. This paper answers both of the aforementioned questions presenting a first fully automatic food intake detection methodology based on wearable non-invasive swallowing sensor and validating it for intra- and inter-subject models on a large database collected from 12 subjects during food intake and resting periods. Another major contribution of this paper is utilization of principal component analysis and smoothing algorithm to improve automatic swallowing detection accuracy for intra- and inter-subject models.

6.4. Methodology

6.4.1. Human Study

Automatic food intake detection methodology proposed in this paper was validated on a dataset that is a subset of the data collected during the human study reported in [Sazonov et al. 2008]. Short summary of the aspects of the original dataset that are relevant to current study is presented below.

Original subject population included 21 generally healthy volunteers with different degrees of adiposity. Each subject participated in four separate visits scheduled for different days. Each visit consisted of a 20-minute resting period, followed by a meal, followed by another 20-minute resting period. Meals consisted of a fixed sequence of food items selected to represent different physical properties of the food. Subjects ate in silence during half of the meals and were involved in a dialogue during the other half to evaluate the impact of a meal-time conversation on the accuracy of swallowing detection. Additionally, a mix of background noise was used during a half of the visits to simulate realistic environments where people may be eating. Subjects were monitored by a multi-modal sensor system which included an IASUS NT (IASUS Concepts Ltd) throat microphone located over laryngopharynx. The recordings were manually scored to mark the boundaries of food intake periods and each swallowing instance. The evaluation of inter-rater reliability showed high reliability of manual swallowing scores with average intra-class correlation coefficient of 0.98

To our knowledge this is the largest dataset collected to date in a study of ingestive behavior monitoring based of data from wearable non-invasive sensors. It is also the most

complicated one with inclusion of variety of sound artifacts and background noises of various origins, variety of food items and human subjects with different degrees of adiposity to create experimental conditions resembling those of free-living food consumption. A complete review on the original dataset including details of the protocol, hardware, software and reliability of the manual scoring procedure can be found in [Sazonov et al. 2008].

Out of total 84 originally collected visits, 4 visits collected from one subject were used for initial calibration of the multi-modal monitoring system and therefore discarded from further studies and other 10 visits had partially incomplete data due to different operator's errors committed during the data collection process. Even though most of these errors were minor these visits were discarded from the dataset. From the remaining 70 complete visits only 12 out of 20 subjects had complete data for all four visits. These 12 subjects comprise the dataset used to validate the methodology proposed in this paper. This derived dataset is large (44.1 hours with a total of 7305 swallows) and as complicated as original one since it includes the same variety of data for population with similar average degree of adiposity. Namely, the average BMI for the derived dataset is 29.2±6.9 compared to 29±6.4 of the original dataset. Average intra-visit swallowing detection accuracy calculated for the derived dataset is 96.7% (per-epoch) and 85.1% (per-swallow) compared to 96.8% and 84.7% respectively obtained for the original dataset which is another indication that derived dataset is a representative subset of the original one and can be used as such in this study.

6.4.2. Automatic Detection of Swallowing Instances

Methodology for automatic detection of swallowing instances by acoustical means presented in this paper is based on the one proposed in [Sazonov et al. 2010] with two major improvements. Summary of the original methodology and description of proposed improvements are presented below.

The original methodology was based on mel-scale Fourier Spectrum (msFS) for time-frequency representation and support vector machines (SVM) for automatic recognition of characteristic sound of swallowing [Sazonov et al. 2010]. First, the sound stream was split into a series of overlapping epochs and mel-scale Fourier transform were applied to each epoch. Second, resulting epoch feature vectors were merged for a number of adjacent epochs to produce a time-lagged feature vectors accounting for time-varying structure of a swallow. These time-lagged vectors were used as inputs for training and validation of SVM classifier using the Gaussian radial basis kernel function. Near-optimal values for the following parameters: epoch duration of 1.5 s, epoch step size of 0.2 s, eighth msFS decomposition level, number of lags equal to 1, SVM misclassification penalty parameter equal to 10 and Gaussian kernel width parameter equal to 0.05 were determined using a grid search procedure in original work and used in current study. The methodology was tested on the original dataset containing 70 visits from 20 subjects [Sazonov et al. 2008]. Each visit was divided into 55 equal segments with average duration of 1 minute. Three-fold cross-validation was performed with two folds or two of every three consecutive segments used for training and one fold used for validation at each step. The average accuracy of 96.8% for epochs and 84.7% for swallowing instances was obtained for such intra-visit models.

In this paper we propose and test two major improvements to the original methodology presented in [Sazonov et al. 2010]. To improve the performance of automatic detection of swallowing instances we propose a preprocessing of msFS features using principal component analysis (PCA) and postprocessing of the automatic predicted epochs using a smoothing algorithm.

The combination of two machine learning algorithms: supervised SVM and unsupervised PCA is widely used in biomedical engineering [Jin et al. 2007, Rong et al. 2008]. PCA is a multivariate non-parametric statistical technique that being applied to a number of possibly correlated variables allows to reveal the internal structure of the data in a way that best explains its variance and to transform the data into a new set of orthogonal variables called principal components which are linear combinations of the original variables. The first principal component accounts for as much of the variance in the original data as possible, and each succeeding component accounts for as much of the remaining variance as possible. Detailed description of the PCA mechanism based on calculation of the eigenvalue decomposition of a data covariance matrix is out of the scope of this paper and can be found in [Jolliffe 2002]. Measuring variance along each principal component provides information on the relative importance of each component. Therefore, PCA is often used for dimensionality reduction of feature vectors with smaller number of principal components being used compared to the original feature vector dimension. Since kernel methods like SVM are tolerant to high dimensionality of features we can use SVM with maximal number of principal components not losing any data from the original dataset.

A postprocessing smoothing algorithm is proposed to refine the automatic epoch score produced by the SVM. Binary class labels assigned to epochs for training and prediction of automatic score by SVM were produced in the following way: if any part of the epoch belonged to a swallow mark in the manual score the epoch label was marked as '1' (swallow epoch), otherwise it was marked as '-1' (non-swallow epoch). Such predicted automatic score represents accuracy of the classifier on the epoch level. To evaluate the accuracy of the classifier in detection of swallowing events manual and automatic label scores were used to identify all the situations were either score indicated presence of a swallow and calculating the number of true positive, false positive, false negative and true negative detections in terms of swallowing events. These detections were later used to calculate sensitivity, specificity and prevalence to further calculate overall accuracy of swallowing detection as a weighted average of sensitivity and specificity [Alberg et al. 2004]. Refinement of the automatic epoch score was performed in the two steps: first, labels for short segments of up to a predefined number of epochs in duration that were automatically marked as '-1' (non-swallow epochs) but were surrounded by epochs marked as '1' (swallow epochs) on both sides were reset to '1'. This postprocessing step was needed to correct the situations in which a single swallow of more than two epochs may be split into several parts by a misclassified epoch. Second, labels for short segments of up to a predefined number of epochs in duration that were automatically marked as '1' but were surrounded by epochs marked as '-1' on both sides were reset to '-1' fixing the cases where accidental epochs were incorrectly classified as swallows. Optimal values of predefined numbers of epochs for swallow and non-swallow gaps to be reset by the

smoothing algorithm were obtained using grid search and were equal to 5 and 0 respectively.

The results of testing of both the original and improved swallowing detection techniques as a part of a food intake detection methodology in intra- and inter-subject models are presented below.

6.4.3. Automatic Detection of Food Intake

Methodology for automatic detection of food intake based on swallowing presented in this paper stems from the one proposed in [Sazonov et al. 2009].

Food intake prediction models assign binary labels ‘intake’ or ‘no intake’ to time windows of predefined length based on average instantaneous swallowing frequency (ISF) calculated for current window. ISF stands for the inverted difference in time between each two consecutive swallows and as an instant frequency is expressed in swallows per minute. Higher ISF value indicates shorter time between two consecutive swallows.

Selection of the window size defines the time resolution of intake detection. As higher frequencies of swallowing indicate the presence of food ingestion the window should be long enough to detect an increase in the frequency of swallowing. At the same time it should be short enough to detect such short food consumption events as snacking. In [Sazonov et al. 2009] we estimated the optimal trade-off between detection accuracy and time resolution to be a time window length of 30 s. The same window length was used in this study.

Detection of food intake using swallowing frequency as a predictor was performed in the following way using floating average prediction model: first, a decision threshold is calculated as a product of the average ISF for the training set multiplied by a scaling factor. In this way a decision threshold is a function of the average ISF. Prediction is built for each time window with food intake being detected if average ISF for current window is higher than the decision threshold and no food intake being detected otherwise. Training is repeated for a range of scaling factors and optimal scaling factor is selected based on the highest accuracy achieved with the training set prediction and further used for intake detection on the validation set. Prediction on the validation set is built in the same way as on the training set using the optimal scaling factor to obtain the decision threshold. Complete details on this approach can be found in [Sazonov et al. 2009] where similar model was proposed and validated with an average accuracy of 87% obtained on the manual scores from the dataset described in [Sazonov et al. 2008]. Validation on automatic scores of detected swallowing instances as a part of automatic food intake detection for intra- and inter-subject models are presented below.

6.5. Experimental results

6.5.1. Intra-subject Food Intake Detection Models

Subject-specific intra-subject food intake detection models were built separately for each subject in such a way that these models were completely independent from each other as only the data obtained from current subject was used to build and validate the model.

Four-fold cross-validation was used with four folds being the four visits of current subject. At each cross-validation step both swallowing detection and food intake

detection were performed with three visits of a subject used for training and the remaining one used for validation. Namely, each cross-validation step consisted of:

- Swallowing detection stage: acoustical sensor data and manual swallowing scores for three training visits were used to produce the automatic swallowing score for the validation visit.

- Food-intake detection stage: manual swallowing and food intake scores for three training visits were used to obtain an optimal scaling factor that was further used for intake detection on the validation visit using the automatic swallowing score produced at the previous stage.

- Accuracy assessment: manual food intake score for the validation visit was used to calculate the accuracy of food intake detection.

The results of automatic detection of swallowing and food intake for intra-subject models with and without preprocessing with PCA and postprocessing with smoothing algorithm are presented in Table 3. Per subject and average receiver operating characteristic (ROC) for intra-subject food intake detection models were built with sensitivity and specificity values obtained using a range of scaling factors to create a range of intake predictions on the validation set as compared to using a single optimal scaling factor. These ROC curves are presented in Fig. 13. Finally, distributions of intra-subject swallowing and food intake detection accuracies versus the subject's BMI and corresponding linear fits of the data are presented in figure 14. Figures 13 and 14 are built for the case of the highest average accuracy obtained for intra-subject model highlighted with bold in Table 3.

Table 3: Effects of preprocessing (PCA) and postprocessing (smoothing algorithm) on average accuracy for intra-subject model

Intra-subject model	Average accuracy [Sensitivity, Specificity] (%)			
	Baseline	Baseline + preprocessing	Baseline + postprocessing	Baseline + preprocessing + postprocessing
Per-epoch swallowing detection	95.1 [39.3, 98.7]	95.7 [42.6, 99.2]	95 [40.7, 98.5]	**95.7 [44, 99]**
Per-swallow swallowing detection	74.5 [66.2, 81.7]	79.2 [72.5, 84]	75.9 [64.9, 84.8]	**80.4 [71.3, 87]**
Food intake detection	76.3 [68.3, 79]	75.7 [72, 77.2]	77.7 [67.7, 81.2]	**80.1 [72.1, 83.7]**

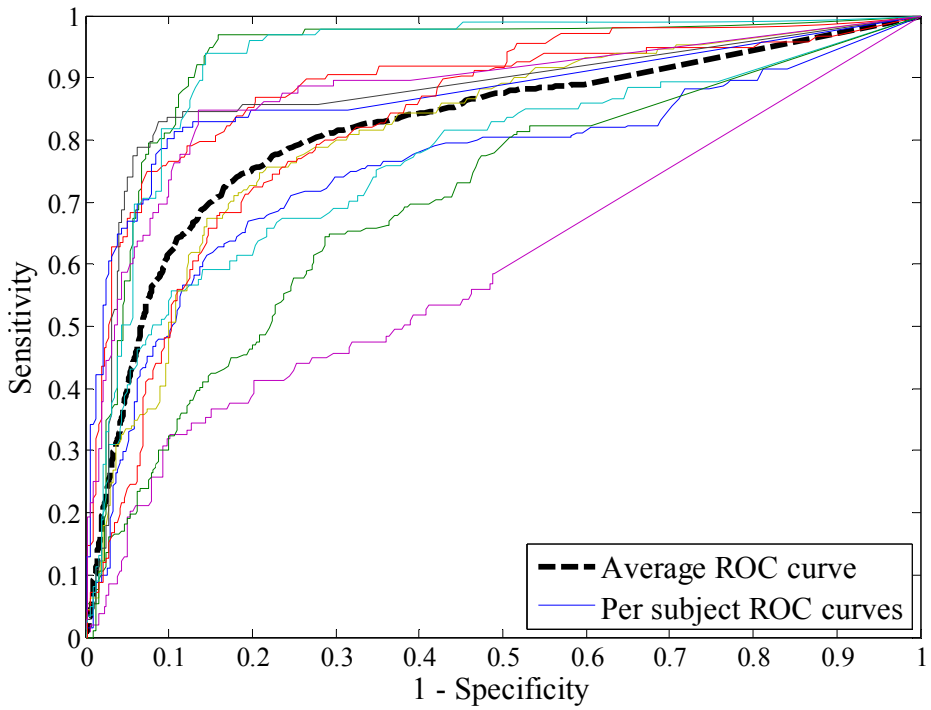


Figure 13: ROC curves for intra-subject model: per subject and average for all subjects.

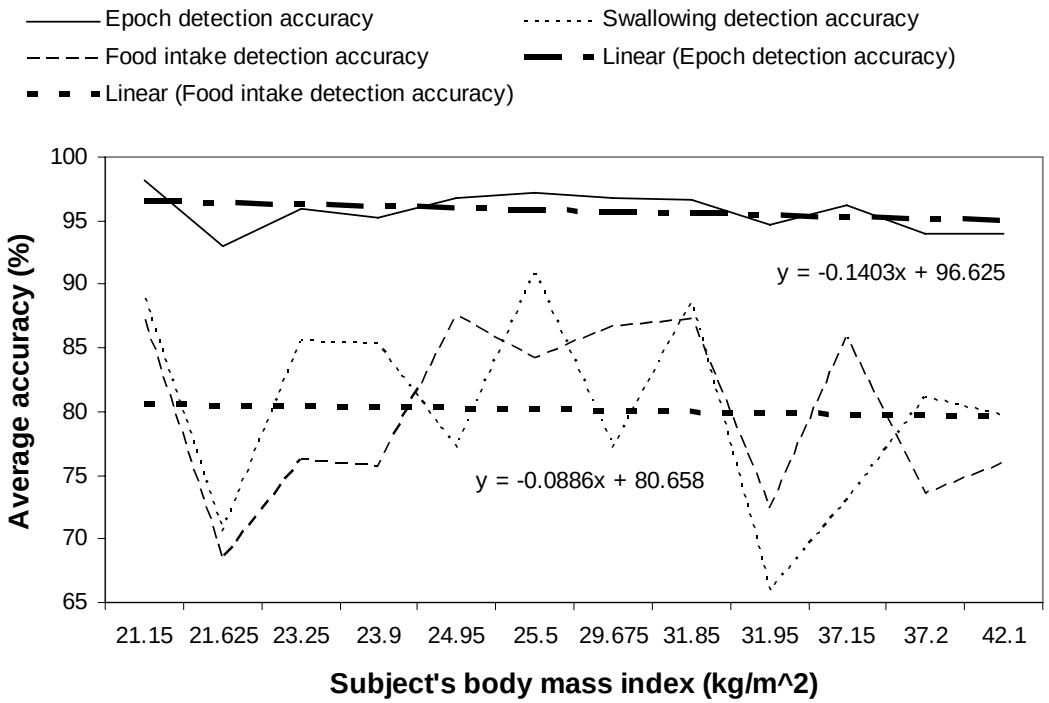


Figure 14: Distribution of average accuracy in per-epoch and per-swallow swallowing detection versus subject's BMI for intra-subject model.

6.5.2. Inter-subject Food Intake Detection Models

Non-personalized inter-subject food intake detection models were built separately for each subject in such a way that the training and validation sets belonged to non-intersecting subsets of data.

Twelve-fold cross-validation was used with each fold representing all four visits of a certain subject. At each cross-validation step all the data from eleven subjects was used for training and all the data from the remaining one subject was used for validation. Namely, each cross-validation step consisted of:

- Swallowing detection stage: acoustical sensor data and manual swallowing scores for data from eleven subjects were used to produce the automatic swallowing scores for data from the remaining subject.

- Food-intake detection stage: manual swallowing and food intake scores for data from eleven subjects were used to obtain an optimal scaling factor that was further used for intake detection on the data from the remaining subject using the automatic swallowing scores produced at the previous stage.

- Accuracy assessment: manual food intake scores for data from the validation subject were used to calculate the accuracy of food intake detection.

The results of automatic detection of swallowing and food intake for inter-subject models with and without preprocessing with PCA and postprocessing with smoothing algorithm are presented in Table 4. Per subject and average receiver operating characteristic (ROC) for inter-subject food intake detection were built in the same way as for the intra-subject models. These ROC curves are presented in figure 15. Distributions of inter-subject swallowing and food intake detection accuracies versus the subject's BMI and corresponding linear fits of the data are presented in figure 16. Figures 15 and 16 are built for the case of the highest average accuracy obtained for inter-subject model highlighted with bold in Table 4.

Table 4: Effects of preprocessing (PCA) and postprocessing (smoothing algorithm) on average accuracy for inter-subject model

Inter-subject model	Average accuracy [Sensitivity, Specificity] (%)			
	Baseline	Baseline + preprocessing	Baseline + postprocessing	Baseline + preprocessing + postprocessing
Per-epoch swallowing detection	91.5 [36.3, 95.2]	93.5 [25, 98]	**90.9 [39, 94.3]**	93.3 [26.5, 97.8]
Per-swallow swallowing detection	64 [62.1, 69.6]	65.3 [52.1, 72.9]	**66.4 [60, 73.1]**	66.7 [51.5, 75.6]
Food intake detection	70.4 [74.2, 74.9]	59.9 [57.3, 58.7]	**70.9 [75.5, 77]**	61.8 [58.2, 59.4]

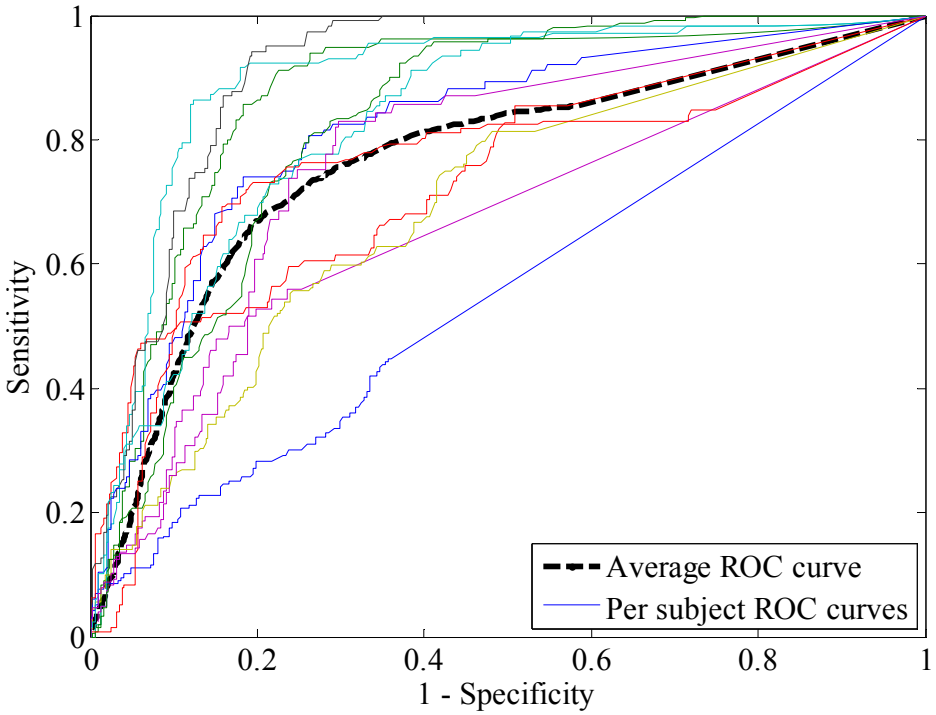


Figure 15: ROC curves for inter-subject model: per subject and average for all subjects.

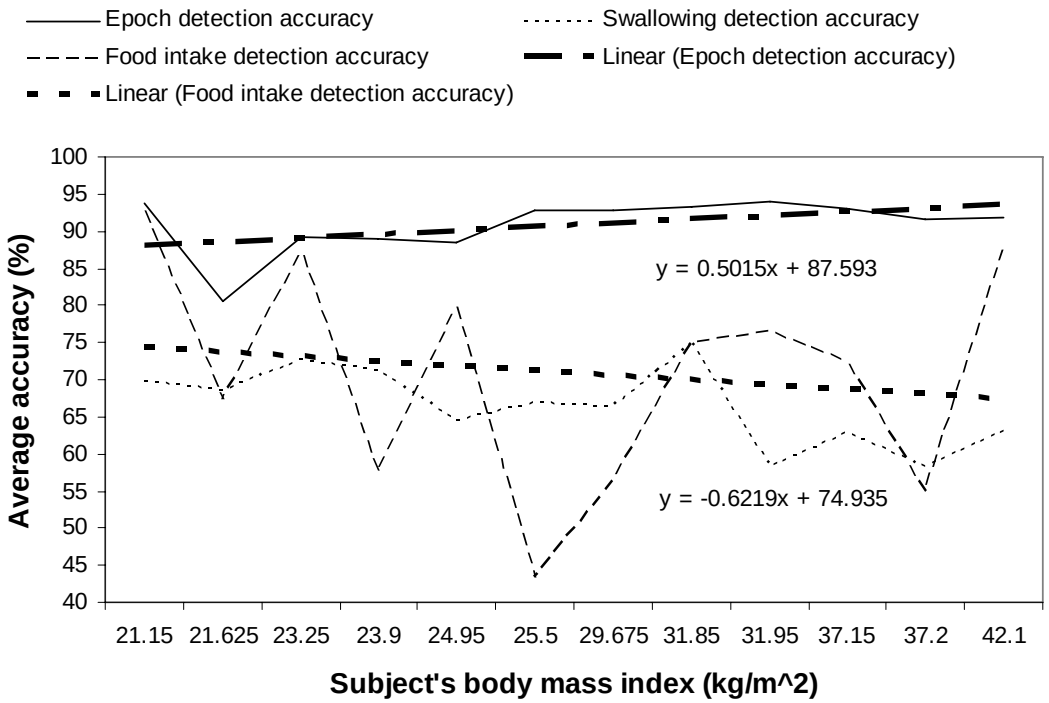


Figure 16: Distribution of average accuracy in per-epoch and per-swallow swallowing detection versus subject's BMI for inter-subject model.

6.6. Discussion

As can be seen from Tables 3 and 4 the highest average food intake detection accuracies of 80.1% and 70.9% were obtained for intra- and inter-subject models respectively. Figures 13 and 15 show a reasonable tradeoff between sensitivity and specificity for both models. We can also see that for each model there are no more than four subjects demonstrating a receiver operating characteristic that is significantly worse than the average with a majority of subjects demonstrating a comparable of better characteristic. The reasons for such a significant dispersion of results for different subjects are yet to be determined. One of the valid hypotheses is dependency of intake detection accuracy on subject's BMI. However, linear fits for distributions of average intake detection accuracy versus subject's BMI presented in figures 14 and 16 suggest that this hypothesis is likely

to be false. Even though more data is needed to achieve higher statistical significance these linear fits also suggest that the proposed food intake detection approach could be used for monitoring of obese individuals.

It can be seen from Tables 3 and 4 that effect of feature preprocessing using PCA was different for intra- and inter-subject models. In the former case it allowed to improve the detection accuracy while in the latter it didn't. We believe that this difference can be attributed to the difference in PCA application for two models. This difference stems from computational burden of covariance matrix calculation needed to compute the principal components. Dimensionality of the covariance matrix is A by A where A is the number of observations or, in our case, epochs in the dataset. Covariance matrix is calculated for the training set and is used to calculate the matrix of eigenvalues which is further used to project the validation data onto the new orthogonal basis. For intra-subject models such training set is limited to three visits from a particular subject and covariance matrix for it can be calculated directly even though this is the biggest feasible number of visits as was determined empirically. For intra-subject models the training set includes a total of 44 visits from eleven subjects. Such more than fourteen times increase in the size of the training set makes calculation of a complete covariance matrix infeasible. Therefore, subset of 3 visits was selected randomly from the original training set and PCA training was performed on this subset only. Representativeness of such a small subset is very limited, so a different scheme of PCA application for inter-subject models allowing a better representation of the training set data in covariance matrix could potentially result in an increase in intake detection accuracy similar to the one obtained for intra-subject models.

As different from the PCA based preprocessing, postprocessing with a smoothing algorithm allowed to improve the detection accuracy for both models as can be seen in tables 3 and 4. Optimal values of thresholds found in grid search suggest that situations where not all the epochs belonging to a single swallow are classified correctly thus creating additional false positive swallow detection are a source of swallowing detection errors. We can also see that in half of the cases food intake detection offers some tolerance to errors in automatic swallowing detection resulting in a higher average accuracy. This is due to food intake detection being based of time intervals containing multiple swallows which allows some compensation in swallowing detection errors. Further investigation is needed for more precise evaluation of the effect of errors in automatic swallowing score on accuracy of food intake detection.

Finally, from tables 3 and 5 and section 6.4.1 we can see that even with the proposed improvements the highest per-epoch and per-swallow swallowing detection accuracies for intra- and inter-subject models are lower than the ones obtained with training and validation performed on each visit separately as in [Sazonov et al. 2010]. Several factors could be contributing to this effect besides the intra- and inter-subject variability of swallowing sounds including inconsistencies of positioning and fixation of the sound sensor for different visits and subjects and further investigation of it is needed.

While the direct comparison of the swallowing based automatic food intake detection approach proposed in this paper with intake gesture and chewing based approaches proposed in [Junker et al. 2008] and [Nishimura and Kuroda 2008] respectively can not be drawn in the strictest sense we can compare the validation procedures. In [Nishimura

and Kuroda 2008] very limited detail is provided on how many human subjects participated in the study, whether intra- or inter-subject model results are reported, how the data was divided into training and validation sets, what validation technique was used, etc. Absence of this critical information on validation procedure does not allow us to interpret the reported high average detection accuracy of 98.7% and makes the proposed approach of limited practical application. In [Junker et al. 2008] comparable results (average recall and precision of 0.78 and 0.77 respectively) are reported to the ones obtained in this study for subject-specific intra-subject model validated on a smaller dataset with a total of 784 intake gestures compared to a total of 7305 swallows in our study. No results for non-personalized inter-subject model were reported in [Junker et al. 2008] compared to the ones reported in this paper. Therefore, we can conclude that validation procedure presented in this paper is the most complete and was performed on the largest dataset. Since the proposed swallowing based approach is free of limitations inherent in intake gesture and chewing based approaches (outlined in section 6.3) we conclude that swallowing sensor may be a most promising option for creation of a single sensor MIB device.

Even though both intra- and inter-subject models can be implemented in a wearable food intake monitor an important advantage of non-personalized inter-subject model is that it can be applied to any subject without any prior training on the data from this subject. Implementation of the intra-subject model would require device calibration for each particular subject prior to beginning of monitoring. During this calibration subject would have to manually indicate all his swallowing instances with, for example, pushing the handheld push-button device during food intake and resting periods. Minimal duration of

resting period and minimal amount of food items necessary for obtaining sufficient training data may be required. These indicated swallows will be used as manual swallowing scores for training data. Such button-based calibration resembles, for example, fixed volume bag based calibration of the plethysmograph used in LifeShirt (VivoMetrics Inc) [Wilhelm et al. 2003].

As practical applicability of the proposed food intake detection approach at this point of time is limited by achieved accuracy to intra-subject models with a separate study needed to further improve inter-subject models we can outline the following directions of future work. First, a new human study needs to be conducted to confirm the effectiveness of the proposed approach in long-term (24 hours) and virtually unrestricted free-living conditions. Second, using sensor fusion for food intake detection based on combined data from different types of sensors may be advantageous. For example, prediction model proposed in [Sazonov et al. 2009] was able to detect periods of food intake with average accuracy of 87% achieved for 30s windows based on swallows only while a similar model based on chews and swallows was able to achieve average accuracy of 95.5%. On the other hand, using just one sensor may significantly simplify operation of the MIB device and reduce its cost.

6.7. Conclusion

First fully automatic food intake detection methodology based on wearable non-invasive swallowing sensor is proposed and validated on the large dataset. Utilization of principal component analysis and smoothing algorithm allowed to obtain average accuracies of >80% and >70% for intra-subject and inter-subject models that can be implemented

directly in a wearable device that automatically monitors ingestive behavior in humans. Such a device can potentially be used in free-living conditions improving our understanding of eating behaviors associated with obesity and other eating disorders and providing the real-time biofeedback to individuals. With the rapid progression of computing power available in modern ubiquitous platforms this device may potentially find numerous applications in research, clinical nutrition and self-monitoring of food intake by general population.

7. Overall Conclusions and Future Work

This chapter summarizes the unique contributions presented in this dissertation and an overview of future work.

7.1. Feasibility of Artifact Elimination with an Acoustical Swallowing Detection Method

- Novel swallowing sound recognition methodology based on the limited receptive area (LIRA) neural classifier and time-frequency decomposition was proposed and tested in recognition of four classes of sounds that correspond to swallowing sounds, talking, head movements and outlier sounds with two different algorithms of time-frequency decomposition, short-time Fourier transform (STFT) and continuous wavelet transform (CWT) showing feasibility of elimination of such artifacts with an acoustical swallowing detection method and efficiency and reliability of the proposed method.

7.2. Automatic Detection of Swallowing Events by Acoustical Means

- Two novel automatic acoustical swallowing detection methods based on combination of mel-scale Fourier Spectrum (msFS) or Wavelet Packet Decomposition (WPD) and Support Vector Machines were proposed and tested on the data collected from 20 human subjects with 35% of the subjects being obese using a multi-modal data collection system designed for non-invasive monitoring of chewing and swallowing.

- The total duration of data used for training and validation was 64.5 hours including 9,966 swallows. To our knowledge this is the largest dataset of its kind collected to date. Being designed to resemble food consumption in free living our dataset includes sound artifacts originating from chewing of food of different textures, talking, head movements, occasional intrinsic sounds, and background noise of various origins making it also the most complicated dataset collected to date.

- Effects of epoch size, level of decomposition and lagging on classification accuracy were studied yielding near-optimal parameter values and showing that msFS decomposition based methodology outperforms the WPD based one. High average weighted epoch recognition accuracy 96.8% was obtained for intra-visit individual models resulting in 84.7% average weighted accuracy in detection of swallowing events.

- Study of impact of such sound artifacts as food intake, talking and background noise on accuracy of swallowing detection suggested robustness of the proposed methodology to such events as well as its ability to accurately separate swallowing sounds from sound artifacts that originate in respiration, talking, head movements, food ingestion, and ambient noise.

- The proposed methodology was also demonstrated to work equally well for both obese and non-obese subjects.

7.3. Automatic Food Intake Detection Based on Swallowing Sounds

- Novel fully automatic food intake detection methodology based on wearable non-invasive swallowing sensor was proposed and tested on the data collected from 12 subjects during food intake and resting periods. This large dataset of 44.1 hours with a total of 7305 swallows is derived from the original dataset used for validation of our automatic swallowing detection methodology is as complicated as the original one including the same variety of data for population with similar average degree of adiposity.

- To improve the performance of automatic detection of swallowing instances preprocessing of msFS features using principal component analysis (PCA) and postprocessing of the automatic predicted epochs using a smoothing algorithm were proposed.

- Average food intake detection accuracies of >80% and >70% were obtained for intra-subject and inter-subject models correspondingly with detection accuracies being comparable for obese and lean subjects.

- These models can be implemented directly in a wearable device that automatically monitors ingestive behavior in humans. Such a device can potentially be used in free-living conditions improving our understanding of eating behaviors associated with obesity and other eating disorders and providing the real-time biofeedback to individuals.

- To the best of our knowledge this is the first attempt of fully automatic detection of food intake based on the data from a wearable non-invasive swallowing sensor.

- Limitations inherent in two previously proposed approaches based on intake gestures and chewing sensors suggest that swallowing sensor may be a most promising option for creation of a single sensor biofeedback device.

7.4. Future Work

Results obtained for automatic swallowing and food intake detection presented in this dissertation suggest that while intra-subject models may be practically applicable at this point of time a separate study is needed to further improve inter-subject models. Below is an outline of the directions for future work regarding the proposed approach in general followed by discussion of possible improvements to its components.

As a future work on the proposed approach in general, first, a new human study needs to be conducted to confirm the effectiveness of the proposed approach in long-term (24 hours) and virtually unrestricted free-living conditions as opposite to laboratory settings of the current study. For example, a commercially available pocket-size MP3 player/recorder can be used to record approximately a day's worth of data from a subject. The swallowing sound data can captured by a miniature throat microphone similar to the one used in current study. The subject will be instructed to keep a very strict diary of every food consumption event during the day. The beginning of the data collection will take part in laboratory settings and will be similar in structure and duration to data collection visits used in current study. This data will be used for training of swallowing and food intake detection methodologies. After that the subject will keep wearing the

recorder for the rest of the day in free-living conditions and this portion of the data will be used to automatically detect swallowing instances and periods of food intake. Food intake detection accuracy will be assessed based on the subject's diary.

Second, using sensor fusion for food intake detection based on combined data from different types of sensors may be advantageous. For example, prediction model proposed in [Sazonov et al. 2009] was able to detect periods of food intake with average accuracy of 87% achieved for 30s windows based on swallows only while a similar model based on chews and swallows was able to achieve average accuracy of 95.5%. In [Amft and Tröster 2009] a model of intake cycle incorporating the data from intake gesture, chewing and swallowing sensor is proposed. It contains a hierarchical recognition procedure to identify intake cycles that can potentially compensate for individual sensor's errors and help overcome their drawbacks. Such model may have potential for automatic food intake detection but it needs to be verified through testing.

As a future work on the components of proposed approach, results presented in this dissertation show that even with two improvements proposed for automatic swallowing detection methodology in [Makeyev et al. 2010] the highest per-epoch and per-swallow detection accuracies obtained for intra- and inter-subject models are lower than the ones obtained with training and validation performed on each visit separately in [Sazonov et al. 2010]. Several factors could be contributing to this effect besides the intra- and inter-subject variability of swallowing sounds including inconsistencies of positioning and fixation of the sound sensor for different visits and subjects. Further investigation of this effect is needed. Subjective evaluation of inconsistencies of positioning and fixation of

the sound sensor found in data collected for current study can be found in Appendix B. A possible solution would be to modify the currently used microphone removing the arc and placing the sensor on a band with a partially rigid front part similar to the ones used in neck protectors. Placing a sound sensor in the center of such band will ensure that it will be located close to the trachea at constant height and that its positioning will be consistent. The drawback of this modification is that such band is hard to disguise and creation of a socially acceptable monitoring device based on such band is problematic.

Furthermore, throat microphone used for current study has been selected while automatic swallowing and food intake detection methodologies were still in development. Microphone quality was evaluated using recordings of several consecutive swallows with subsequent evaluation of sound quality both subjectively (i.e. by listening to the recording) and objectively (by visualizing in a sound editor and computing the signal-to-noise ratio). Only four microphones were tested located immediately below the laryngeal prominence and the one with the higher sensitivity to swallowing sounds and lower degree of sensitivity to ambient noise was selected. In future more microphones can be tested for automatic swallowing and food intake detection using the proposed methodologies in intra- and inter-subject models with highest detection accuracy being the selection criteria. The choice of microphones to be tested may be narrowed down based on the size, power consumption and sensitivity. New positioning and fixation methods can be tested in the same way. For example, new location on the sternum below the laryngeal prominence has an advantage of being sufficiently close to the source of the sound and the sensor can be easily hidden under large variety of clothing while in case of current location the microphone can only be hidden under certain type of clothing. An

example of potential new fixation method would be holding the microphone against the skin by a Band-Aid type adhesive representing an easily disguisable alternative to currently used band.

Finally, an outline of potential improvements and alternatives to the algorithms comprising the proposed swallowing detection methodology is presented below. Out of all time-frequency decomposition techniques currently widely used in sound recognition the most promising alternative to msFS and WPD seems to be the Hilber-Huang Transform (HHT). Evaluation of HHT in terms of its comparison to other time-frequency decomposition techniques is presented in Appendix C. In future work HHT based features can be evaluated both on linear and mel-scale. Out of all supervised machine learning methods currently widely used in sound recognition the most promising alternative to LIRA and SVM seems to be Hidden Markov Models (HMM). Based on empirical risk minimization principle HMM usually show comparable or lower performance in direct comparison with SVM that are based on structural risk minimization principle controlling not only the empirical risk on the training data set but also the capacity of the decision functions used to obtain the risk value [Justino et al. 2005]. However, several recently proposed HMM-SVM hybridization schemes have shown higher recognition rates compared to either of composing methods [Valstar and Pantic 2007, Krüger et al. 2005]. The same have recently been shown for SVM hybridization with Genetic Algorithms (GA) where GA was used to optimize both feature subset selection and parameters of SVM [Min et al. 2006].

Overall, directions of future work outlined in this chapter should allow to improve current swallowing and food intake detection accuracies for both intra- and inter-subject models further confirming high potential of the approach proposed in this dissertation.

References

Aboofazeli M and Moussavi Z (2004) "Automated classification of swallowing and breath sounds" In: Proceedings of 26th Annual International Conference of the Engineering in Medicine and Biology Society, San Francisco, 3816-3819.

Aboofazeli M and Moussavi Z (2005) "Analysis and Classification of Swallowing Sounds Using Reconstructed Phase Space Features", Proc. IEEE ICASSP, pp. V421-V424.

Aboofazeli M and Moussavi Z (2006) "Automated Extraction of Swallowing Sounds Using a Wavelet-Based Filter," in Proc. 28th Annual International Conference of the Engineering in Medicine and Biology Society, New York, New York, USA, pp. 5607-5610.

Aboofazeli M, Moussavi Z. (2008) "Analysis of swallowing sounds using hidden Markov models." Medical and Biological Engineering and Computing. 46(4):307–314.

Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M (2004) "The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests" J. Gen. Intern. Med. 19(5 Pt 1):460-5.

Amft O, Tröster G. (2008) "Recognition of dietary activity events using on-body sensors. Artificial Intelligence in Medicine." 42(2):121–136.

Amft O, Tröster G. (2009) "On-body Sensing Solutions for automatic dietary monitoring." IEEE Pervasive Computing.8(2):62–70.

Ainslie PN, Reilly T, Westerterp KR (2003) "Estimating human energy expenditure: a review of techniques with particular reference to doubly labeled water." Sports Medicine. 33(9):683–698.

Addison PS (2002) "The illustrated wavelet transform handbook" Institute of Physics Publishing, Bristol.

Baidyk T, Kussul E, Makeyev O, Caballero A, Ruiz L, Carrera G, Velasco G (2004) "Flat image recognition in the process of microdevice assembly". Pattern. Recogn. Lett. 25:107-118.

Buckheit JB, Donoho DL (1995) "Wavelab and reproducible research," in "Wavelets and Statistics" (A. Antoniadis and G. Oppenheim, Eds.) Springer-Verlag, New York.

Carmelli D, Zhang H, and Swan GE (1997) "Obesity and 33-year follow-up for coronary heart disease and cancer mortality." Epidemiology,1997. 8(4):378-83.

Champagne CM, Bray GA, Kurtz AA, Monteiro JB, Tucker E, Volaufova J, Delany JP (2002) "Energy intake and energy expenditure: a controlled study comparing dietitians and non-dietitians." J Am Diet Assoc. 2002 Oct;102(10):1428-32.

Chang CC and Lin CJ, (2001) LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cristianini N and Shawe-Taylor J (2000) "An Introduction to Support Vector Machines and other kernel-based learning methods" Cambridge University Press.

Das A., Reddy NP, Narayanan J, (2000) "Hybrid fuzzy-neural committee networks for recognition of swallow acceleration signals." Comput. Meth. Prog. Bio. 64, 87-99.

De Belie N, Sivertsvik M, De Baerdemaeker J. (2003) "Differences in chewing sounds of dry-crisp snacks by multivariate data analysis." *Journal of Sound and Vibration*. 266(3):625–643.

De Castro JM. (1994) "Methodology, correlational analysis, and interpretation of diet diary records of the food and fluid intake of free-living humans." Appetite. 23(2):179.

Eckel RH and Krauss RM (1998) "American Heart Association call to action: obesity as a major risk factor for coronary heart disease." AHA Nutrition Committee. Circulation, 97(21):2099-100.

Ertekin C, Aydogdu I, Secil Y, et al. (2002) “Oropharyngeal swallowing in craniocervical dystonia.” British Medical Journal.73(4):406.

Flatt JP (1996) “Substrate utilization and obesity”, Diabetes Rev 4:433-449.

Fu S, Muralikrishnan B, Raja J (2003) “Engineering surface analysis with different wavelet bases.” Journal of Manufacturing Science and Engineering. 125:844.

Haines PS, Hama MY, Guilkey DK, Popkin BM. (2003) “Weekend eating in the United States is linked with greater energy, fat, and alcohol intake.” Obesity research. 11(8):945–949.

Hill JO, Wyatt HR, Reed GW, Peters JC (2003) “Obesity and the environment: Where do we go from here?” Science.299(5608):853.

Huang NE, Shen Z, Long SR, Wu MC, Shih EH, Zheng Q, Tung CC and Liu HH, (1998) “The Empirical Mode Decomposition Method and the Hilbert Spectrum for Non-stationary Time Series Analysis” Proc. Roy. Soc. London, A454:903-995.

James WP (1998) “What are the health risks? The medical consequences of obesity and its health risks.” Exp Clin Endocrinol Diabetes, 1998. 106(Suppl 2):1-6.

Jin J, Wang X, Wang B (2007) “Classification of direction perception EEG based on PCA-SVM” Proc. Third Int. Conf. on Natural Computation ICNC’2007 (Haikou, China, 24-27 August 2007) 116-20.

Jolliffe I (2002) “Principal Component Analysis”, second edition, (New York: Springer).

Junker H, Amft O, Lukowicz P, Tröster G (2008) “Gesture spotting with body-worn inertial sensors to detect user activities” Pattern Recogn. 41(6):2010-24.

Justino EJ, Bortolozzi F, Sabourin R (2005) “A comparison of SVM and HMM classifiers in the off-line signature verification” Pattern Recognition Letters 26(9):1377–1385.

Kaczkowski CH, Jones PJ, Feng J, Bayley HS. (2000) "Four-day multimedia diet records underestimate energy needs in middle-aged and elderly women as determined by doubly-labeled water." Journal of Nutrition. 130(4):802.

Kant AK, Ballard-Barbash R, Schatzkin A. (1995) "Evening eating and its relation to self-reported body weight and nutrient intake in women, CSFII 1985-86." Journal of the American College of Nutrition.14(4):358.

Krüger SE, Schafföner M, Katz M, Andelic E, Wendemuth A (2005) "Speech recognition with support vector machines in a hybrid system." In: Ninth European Conference on Speech Communication and Technology INTERSPEECH-2005 (Lisbon, Portugal) 993-996.

Kussul E, Baidyk T (2004). "Improved method of handwritten digit recognition tested on MNIST database". Image. Vision. Comput. 22:971-981.

Kussul E, Baidyk T, Kussul M (2004). "Neural network system for face recognition." In: Proceedings of IEEE International Symposium on Circuits and Systems ISCAS'2004, Vancouver, 768-771.

Kussul E, Baidyk T, Wunsch D, Makeyev O, Martín A (2006) "Permutation coding technique for image recognition systems." IEEE Trans. Neural Networks 17:1566-1579.

Lazareck L and Moussavi Z (2002) "Automated algorithm for swallowing sound detection," in Proc. Canadian Med. and Biol. Eng. Conf.

Lazareck L and Moussavi Z (2004) "Classification of Normal and dysphagic Swallowing Sounds by Acoustical Means", Journal of IEEE, Trans. Biomed. Eng., Vol. 51, NO. 12, pp. 2103-2112.

Lear CS, Flanagan JB and Moorrees CF (1965) "The frequency of deglutition in man" Arch. Oral. Biol. 10:83-100.

Livingstone MBE and Black AE (2003) "Markers of the validity of reported energy intake" J. Nutr. 2003, 133:895–920.

Liu Y, Zhang C, Peng Y (2006) "Neural Classification of Lung Sounds Using Wavelet Packet Coefficients Energy" PRICAI 2006: Trends in Artificial Intelligence, Springer, Berlin p.278-287.

Makeyev O, Sazonov E, Schuckers S, Melanson E and Neuman M (2007a) "Limited receptive area neural classifier for recognition of swallowing sounds using short-time Fourier transform" Proc. Int. Joint Conf. on Neural Networks IJCNN'2007 (Orlando, USA) 1417.1-6.

Makeyev O, Sazonov E, Schuckers S, Lopez-Meyer P, Melanson E and Neuman M (2007b) "Limited receptive area neural classifier for recognition of swallowing sounds using continuous wavelet transform" Proc. of 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society EMBC'2007 (Lyon, France) 3128-31.

Makeyev O, Sazonov E, Baidyk T, Martín A (2008a) "Limited receptive area neural classifier for texture recognition of mechanically treated metal surfaces" Neurocomputing, 71:1413-1421.

Makeyev O, Sazonov E, Schuckers S, Lopez-Meyer P, Baidyk T, Melanson E and Neuman M (2008b) "Recognition of swallowing sounds using time-frequency decomposition and limited receptive area neural classifier" Proc. of 28th SGAI Int. Conf. on Innovative Techniques and Applications of Artificial Intelligence (Cambridge, UK) 33-46.

Makeyev O, Lopez-Meyer P, Schuckers S and Sazonov E (2010) "Automatic food intake detection based on swallowing sounds" Physiological Measurement (In review).

Mertz W, Tsui JC, Judd IT, Reiser S, Hallfrisch J, Morris ER, Steele PD and Lashley E (1991) "What are people really eating? The relation between energy intake derived from estimated diet records and intake determined to maintain body weight" Am. J. Clin. Nutr. 1991 54:291-95.

Meyer D, Leisch F, Hornik K (2003) "The support vector machine under test" Neurocomputing, 55:169-186.

Min SH, Lee J, Han I (2006) "Hybrid genetic algorithms and support vector machines for bankruptcy prediction." Expert Systems with Applications. 31(3):652–660.

Moddemeijer R. (1989) “On estimation of entropy and mutual information of continuous distributions.” Signal Processing. 16(3):233–248.

Montgomery DC (2004) “Design and analysis of experiments” Wiley, Hoboken.

Neumark-Sztainer D, Wall M, Guo J, Story M, Haines J, Eisenberg M (2006) “Obesity, disordered eating, and eating disorders in a longitudinal study of adolescents: how do dieters fare 5 years later?” J. Am. Diet. Assoc. 106(4):559-68.

Nishimura J, Kuroda T (2008) “Eating habits monitoring using wireless wearable in-ear microphone” Proc. Third Int. Symp. on Wireless Pervasive Computing ISWPC’2008 (Santorini, Greece, 7-9 May 2008) 130-2.

Olshansky SJ, Passaro DJ, Hershow RC, Layden J, Carnes BA, Brody J, Hayflick L, Butler RN, Allison DB, Ludwig DS (2005) “A potential decline in life expectancy in the United States in the 21st century” N Engl J Med 2005 Mar 17; 352(11):1138-45.

Palmer JB, Rudin NJ, Lara G, Crompton AW. (1992) “Coordination of mastication and swallowing.” Dysphagia.7(4):187–200.

Pehlivan M, Yuceyar N, Ertekin C, Celebi G, Ertas M, Kalayci T, Aydogdu I (1996) “An electronic device measuring the frequency of spontaneous swallowing: digital phagometer.” Dysphagia.11(4):259-64.

Poppitt SD, Swann D, Black AE and Prentice AM (1998) “Assessment of selective under-reporting of food intake by both obese and non-obese women in a metabolic facility.” Int. J. Obes. 22:303-311.

Prentice AM, Black AE, Murgatroyd PR, Goldberg GR, Coward WA (1989) “Metabolism or appetite: questions of energy balance with particular reference to obesity.” Journal of Human Nutrition and Dietetics 2:965-104.

Rong G, Song-yun X, Xi-na C, Hai-tao Z (2009) “Combined SVM and PCA to recognize the brain function from fMRI images” Proc. Third Int. Conf. in Bioinformatics and Biomedical Engineering ICBBE’2009 (Beijing, China, 11-13 June 2009) 1-3.

Sazonov E, Krishnamurthy V, Makeyev O, Browning R, Hill J, Schutz Y, (2007) "Automatic recognition of postural allocations", in Proc. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBC 2007, Lyon, France, August 23-26, 4993-4996.

Sazonov E, Schuckers S, Lopez-Meyer P, Makeyev O, Sazonova N, Melanson E, Neuman M (2008) "Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior" Physiological Measurement, 29:525-541.

Sazonov E, Schuckers S, Lopez-Meyer P, Makeyev O, Sazonova N, Melanson E, Neuman M (2009a) "Reply to 'Comment on Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior'" Physiological Measurement, 30:L5-L7.

Sazonov E, Schuckers S, Lopez-Meyer P, Makeyev O, Melanson E, Neuman M, Hill J (2009b) "Toward objective monitoring of ingestive behavior in free living population" Obesity, 17:1971–1975.

Sazonov E, Makeyev O, Schuckers S, Lopez-Meyer P, Melanson E, Neuman M (2010) "Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior" IEEE Transactions on Biomedical Engineering, 57:626–633.

Schoeller DA (1988) "Measurement of energy expenditure in free-living humans by using doubly labeled water" J. Nutr. 118(11):1278-89.

Schoeller DA (1995) "Limitations in the assessment of dietary energy intake by self-report Metab." Clin. Exp. 44(2):18-22.

Stellar E, Shrager EE. (1985) "Chews and swallows and the microstructure of eating." American Journal of Clinical Nutrition.42(5):973.

Stunkard AJ. (2002) "Binge eating disorder and the night eating syndrome." Handbook of obesity treatment.107–121.

Subar A, Kipnis V, Troiano R, Midthune D, Schoeller D, Bingham S et al.. (2003) "Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN study" Am J Epidemiol 2003 158:1–13.

Turkoglu I, Arslan A, Ilkay E (2003) "An intelligent system for diagnosis of heart valve diseases with wavelet packet neural networks" Comput. Biol. Med. 33:319-331.

Vaiman M, Nahlieli O. (2009) "Oral vs. pharyngeal dysphagia: surface electromyography randomized study." BMC Ear, Nose and Throat Disorders. 9(1):3.

Valstar M, Pantic M (2007) "Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics" In: Human–Computer Interaction., Springer Berlin / Heidelberg 118-127.

Ward C (1998) "Compulsive Eating: The Struggle to Feed the Hunger Inside." The Rosen Publishing Group.

Weber JL, Reid PM, Greaves KA, DeLany JP, Stanford VA, Going SB, Howell WH, Houtkooper LB (2001) "Validity of self-reported energy intake in lean and obese young women, using two nutrient databases, compared with total energy expenditure assessed by doubly labeled water." Eur J Clin Nutr. 2001 Nov;55(11):940-50.

Wilhelm FH, Roth WT, Sackner MA (2003) "The LifeShirt: an advanced system for ambulatory measurement of respiratory and cardiac function" Behav. Modif. 27(5):671-91.

World Health Organization (2006) Obesity and overweight, Fact sheet N°311 (September 2006) http://www.who.int/mediacentre/factsheets/fs311/en/

Wu GD, Lin CT. (2000) "Word boundary detection with mel-scale frequency bank in noisy environment." IEEE Transactions on Speech and Audio Processing, 8: 541-554.

Wyatt HR and Hill JO (2002) "Let's get serious about promoting physical activity." Am J Clin Nutr, 2002. 75:449-50.

Yon BA, Johnson RK, Harvey-Berino J, Gold BC (2006) "The use of a personal digital assistant for dietary self-monitoring does not improve the validity of self-reports of energy intake" J. Am. Diet. Assoc. 106(8):1256-9.

Youmans SR (2003) "Increasing the objectivity of the clinical dysphagia evaluation: cervical auscultation and tongue function during swallowing" PhD dissertation Florida State University etd-09232003-010436.

Appendix A

Data collection protocol for the conducted human study:

1. Check the equipment that will be used during the data collection:

    a. Check the voltage of the battery to be attached to strain sensor and attach it;

    b. Make sure that the switch on strain sensor is set to “ON”;

    c. Check the correctness of parameters of the Data Acquisition System;

    d. Open the Data Acquisition System (shortcut *DAQ* on the desktop), start recording data, and using the ambient microphone and beeper signal check that all four channels of the preamplifier are operational;

    e. Check that all the microphones that are going to be used during the data collection are operational;

    f. Touch the strain sensor several times and push the button several times, stop recording, check that recorded data is correct;

    g. Turn on the camcorder (shortcut *Motion DV STUDIO LE for DV* on the desktop), capture a short video recording, and make sure that the video is recorded properly;

2. Check the Experiment Form you are going to use. Make sure that fields of the form that correspond to subject ID number, investigator, date and start time of current experiment are filled and this information is accurate.

3. Put on the new pair of exam gloves.

4. Handle the food portion:

    a. Prepare the required portions (weight) of all the food items of the meal to be served. The deviation of up to 5% or 5 grams (whatever is smaller) is allowed for each food item, but try to get as close to the required weight as possible;

    b. For the standard size meal:

        i. Slice of cheese pizza – 80 (76 - 84) grams;

        ii. Can of yogurt – 4 oz (108-118) grams;

        iii. An apple – 80 (76 - 84) grams (slices of the apple without the core);

        iv. A peanut butter sandwich – 37.5 (36 - 39) grams [about 1/2 of the regular peanut butter sandwich: 2 regular slices of bread of 25 grams each and 25 grams of peanut butter];

        v. Water – 200 (195 - 205) grams;

    c. For the large size meal:

i. Slice of cheese pizza – 120 (115 - 125) grams;

ii. Can of yogurt – 6 oz;

iii. An apple – 120 (115 - 125) grams (slices of the apple without the core);

iv. A peanut butter sandwich – 50 (48 - 52) grams [about 3/4 of the regular peanut butter sandwich];

v. Water – 300 (295 - 305) grams;

d. Make sure that all the food items are fresh and satisfy the quality requirements;

e. Make sure that all the food items have appropriate temperature:

i. Take all the food items out of the fridge;

ii. Warm up the pizza in the microwave oven;

5. Take off the exam gloves. You should use the same pair of exam gloves if you'll need to return to handling food later.

6. Input the initial weights of all the food items and water into the corresponding fields of the Experiment Form.

7. Check the Bite Log Sheet that you are going to use. Fill the fields that correspond to the subject ID and session number. Input the initial weights of all the food items and water into the corresponding fields of the Bite Log Sheet.

8. Bring first subject into the lab.

9. Describe the data collection process to the subject:

    a. Point out that all food items are to be consumed unmixed, i.e. separately;

    b. Point out that all the food items are to be consumed completely, i.e. no leftovers;

    c. Point out that the weight of the food item will be measured after each bite;

    d. Point out that water is to be consumed separately from food in the pauses between consumption of different food items;

    e. Explain the sequence of actions to the subject:

        i. 20 min period of inactivity;

        ii. Unbounded time to eat the meal of fixed size plus extra food items at subject's will;

        iii. 20 min period of inactivity;

f. Point out that a part of data collection would involve reading aloud and talking, i.e. the subject would be asked questions not relevant to the research;

g. Point out that the subject is encouraged to abstain from talking unless he is asked to talk. Variety of magazines would be provided to entertain the subject during the inactivity periods;

h. Point out that the subject can't move down his head while reading in order not to obscure the camcorder's perspective of the subject's neck. Subjects are encouraged to read with straight neck, holding the journal in front of the face;

i. Point out that subject's cellular phone is to be turned down during the data collection process.

10. If this is the subject's first visit, have subject read and sign Informed Consent Form, sign the witness field of the form. Write subject's ID on the top of the Informed Consent Form.

11. If this is the subject's first visit, measure the subject's height (cm) and enter the values of subject's height, gender, age and ethnicity into the appropriate fields of the Experiment Form.

12. Measure the subject's weight (kg), waist and hip circumference (inch), calculate the BMI (shortcut BMI Calculator on the desktop), and enter the values of

subject's weight, BMI, waist and hip circumference into the appropriate fields of the Experiment Form.

13. If this is the subject's first visit, fill out the Disbursement Form (name, address, SSN, amount to be paid is $15) and record this in the payment section of the Experiment Form. If this is a nonpayment visit by the subject (they get paid $15 on first visit and $45 after completing 3 more visits) nothing is to be done. If this is the subject's fourth visit, fill out the Disbursement Form (name, address, SSN, amount to be paid is $45) and record this in the payment section of the Experiment Form. Put "Participation in research" in the Explanation field of the Disbursement Form.

14. If this is not subject's fourth visit, set up date and time of the next appointment and input them into the appropriate fields of the Experiment Form.

15. Allow the subject to wash his hands.

16. Put on the new pair of exam gloves.

17. Thoroughly wipe the contact areas (areas that are going to be in contact with the skin of the subject) of all the microphones and strain sensor with the alcohol swabs;

18. Plug the microphones to the following connectors on the preamplifier:

    a. Combined throat-ambient microphone: throat microphone – 1L, ambient microphone – 1R;

b. Throat microphone – 2L;

c. In-ear microphone – 2R;

19. Locate microphones and strain sensor at proper positions and in the following order:

    a. Fully tape the strain sensor to the area immediately below the outer ear;

    b. Locate the combined throat-ambient microphone in the laryngeal area;

    c. Locate the throat microphone on the mastoid bone;

    d. Locate in-ear microphone in the subject's ear;

20. Give subject the button and explain how it should be used.

21. Take off the exam gloves.

22. Make sure that the door to the laboratory is closed. It should remain closed until the end of data collection process.

23. Make sure that camcorder captures the subject from shoulders to the top of the head. Make sure that none of subjects face or neck is cut off or obscured during the period of video recording.

24. Start the data acquisition process:

    a. Fill the "Identity Marker" field of the Data Acquisition System in the form "_x-y-z_" where x is the subject's ID, y is the number of current data

collection session (1 – 4), z is the number of current recording (1 – first inactivity period, 2 – meal, 3 – second inactivity period);

b. Input date and time markers to the Experiment Form;

c. Start video, audio and strain sensor data acquisition;

d. Place the beeper between the subject and the camcorder in such a way that the camcorder would capture the red led light and push the beeper button 4 times, each time holding it for no less than 2 seconds in order to provide a synchronization signal;

25. Set the timer for the first inactivity period.

26. Ask the subject to read silently during the first half of the first inactivity period.

27. Ask the subject to read aloud during the second half of the first inactivity period.

28. When the first inactivity period is over, stop video, audio and strain sensor data acquisition.

29. Serve the meal.

30. If this is the subject's second or fourth visit – turn on the noise recording! Use Panasonic CD stereo system to play noise. Required volume is -40 dB. Ask subject questions to involve him into the conversation;

31. Start the data acquisition process:

a. Fill the “Identity Marker” field of the Data Acquisition System in the form “_x-y-z_” where x is the subject’s ID, y is the number of current data collection session (1 – 4), z is the number of current recording (1 – first inactivity period, 2 – meal, 3 – second inactivity period);

b. Input date and time markers to the Experiment Form;

c. Start video, audio and strain sensor data acquisition;

d. Place the beeper between the subject and the camcorder in such a way that the camcorder would capture the red led light and push the beeper button 4 times, each time holding it for no less than 2 seconds in order to provide a synchronization signal;

32. Tell the subject to start eating. Measure the weight of all the food items and water after each bite and record this data in the corresponding fields of the Bite Log Sheet. Yogurt can is to be placed on the scale by operator and without the spoon! If this is the subject’s first visit, serve a standard size meal, point out to the subject that no background noise is allowed during the data collection. If this is the subject’s second visit, serve a standard size meal, make sure to use background noise and talk to the subject during the data collection. If this is the subject’s third visit, serve a large size meal, point out to the subject that no background noise is allowed during the data collection. If this is the subject’s fourth visit, serve a large size meal, make sure to use background noise and talk to the subject during the data collection.

33. Stop video, audio and strain sensor data acquisition.

34. If the subject still feels hunger, serve the subject extra food items at his will.

35. If this is the subject's first visit, ask the subject if he would have preferred the proposed meal to be larger. Put the subject's answer into the Comments field of the Experiment Form.

36. Start the data acquisition process:

    a. Fill the "Identity Marker" field of the Data Acquisition System in the form "_x-y-z_" where x is the subject's ID, y is the number of current data collection session (1 – 4), z is the number of current recording (1 – first inactivity period, 2 – meal, 3 – second inactivity period);

    b. Input date and time markers to the Experiment Form;

    c. Start video, audio and strain sensor data acquisition;

    d. Place the beeper between the subject and the camcorder in such a way that the camcorder would capture the red led light and push the beeper button 4 times, each time holding it for no less than 2 seconds in order to provide a synchronization signal;

37. Set the timer for the second inactivity period.

38. Ask the subject to read silently during the first half of the second inactivity period.

39. Ask the subject to read aloud during the second half of the second inactivity period.

40. When the second inactivity period is over, stop video, audio and strain sensor data acquisition.

41. Put on the exam gloves. You may use the same pair of gloves that you used locating the sensors on the subject.

42. Take microphones and strain sensor off the subject in the order opposite to the one that was used for putting them on.

43. Inform the subject that the data collection is over and he may leave.

44. Take off the exam gloves.

45. Turn off the strain sensor and camcorder.

46. In case the subject was not able to finish his meal, weight the leftovers of the food items and input the weights into the Experiment Form.

47. Wash the cooking things and clean up the laboratory to the initial state.

48. Fill the end time and comments fields of the Experiment Form; put Bite Log Sheet, Experiment Form, Disbursement Form (if applicable) and Informed Consent Form (if applicable) back to the cabinet.

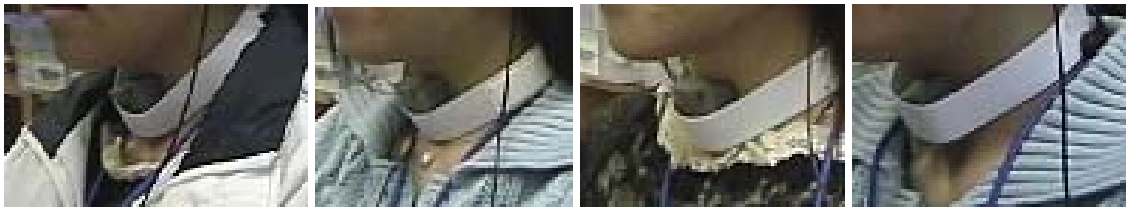
49. Rename of the video files (E:/data/video) in accordance with the names of the corresponding sound files (E:/data/sound).

50. Move all the significant data (3 video, 12 audio and 3 strain sensor data files) obtained during the data collection process to the corresponding folders of the drive D (ULiSATA 5 RAID501).

51. Bring in the next subject.

Appendix B

Even with two improvements proposed for automatic swallowing detection methodology in [Makeyev et al. 2010] the highest per-epoch and per-swallow detection accuracies obtained for intra- and inter-subject models are lower than the ones obtained with training and validation performed on each visit separately in [Sazonov et al. 2010]. Several factors could be contributing to this effect besides the intra- and inter-subject variability of swallowing sounds including inconsistencies of positioning and fixation of the sound sensor for different visits and subjects. In subjective evaluation of inconsistencies of positioning and fixation of the sound sensor found in data collected performed for current study the following hypothesis was proposed: the aforementioned decrease in recognition accuracy is partially due to the variation in positioning and fixation of the throat microphone during the data collection process. The motivation for this hypothesis is the following: positioning of the throat microphone was defined in the data collection protocol used by the operator in the following way: “Locate the combined throat-ambient microphone in the laryngeal area”. Such a guideline defining neither the height at which the microphone should be placed on the neck nor positioning of the microphone relative to the trachea turned out to be insufficient. Despite the fact that during the training data collection visit the operator was instructed on the correct positioning of the microphone, the high importance of positioning accuracy wasn’t emphasized enough since evaluation of the video footage revealed significant inconsistencies in positioning of the microphone for different visits of the same subject (Fig. 17) as well as clearly erroneous positioning (Fig. 18).

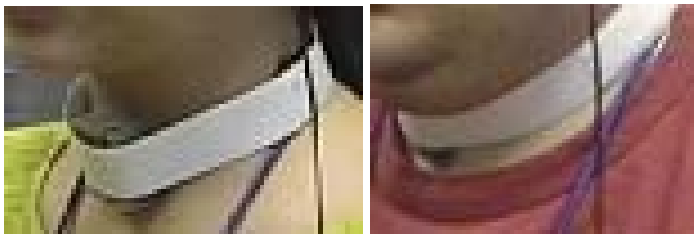


a) b) c) d)

Figure 17: Examples of positioning of the throat microphone for four data collection visits of the same subject.

Figure 18: Example of the incorrect positioning of the throat microphone.

Furthermore, the band with a Velcro fastener that was used during the data collection process to fix the position of the throat microphone resulted to be inefficient as well. First, no guidelines were provided in the protocol on adjustment of the band resulting in a variety of different fittings (Fig. 17). Second, evaluation of the video footage and comments of the operator of the data collection process and the expert rater of the collected data revealed that the band turned out to be inappropriate for most skinny (Fig. 19, a) and obese (Fig. 19, b) subjects, being too large and too small respectively to fix the throat microphone. Exactly the same problem was posed by the holding arc of the throat microphone (Fig. 17, d).



a) b)

Figure 19: Examples of inefficiency of the proposed band in cases of subjects with: a) very low BMI, b) very high BMI.

Third, the operator of the data collection process was changing the positioning of the band on demand from the subject in case the subject was feeling uncomfortable or even taking it off completely in case it was too small. In two cases the part of the subject's data collection visits were performed without the band and in one case all of the subject's data collection visits were performed without the band. We would also like to point out that the overall highest swallowing and food intake detection accuracy was achieved for the human subject who participated in the development of the data collection process and knew what the correct positioning and fixation of the throat microphone should be. Another potential explanation for this fact may be that being the part of the development team this subject paid more attention using the button to indicate swallowing instances which resulted in a more accurate score and therefore more valid training and validation sets of the classifier.

Moreover, evaluation of the video footage revealed that strap was also inefficient in holding the throat microphone in place during the same visit (Fig. 20).

a) b)

Figure 20: Example of inefficiency of the proposed band in holding the throat microphone in place during the same visit: a) beginning of the first session, b) end of the third session.

Therefore, an important step of the future work would be empirical validation of the proposed hypothesis.

Appendix C

Out of all time-frequency decomposition techniques currently widely used in sound recognition including Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), etc. the most promising alternative to msFS and WPD seems to be the Hilber-Huang Transform (HHT). This is due to the advantage of wavelet-based transforms over FFT, DCT or other transforms that express a signal in terms of a sum of sinusoids with different frequencies and amplitudes in tiling of the resolution described in chapter 3 in the example of CWT and STFT. Besides, the underlying assumption of the Fourier-related transforms is stationarity of the data. Furthermore, DWT is a special case of WPD. On the other hand, HHT was developed especially for nonstationary data which is the case of the major part of real physical data. HHT is performed in two steps: first, instantaneous frequencies are calculated based on the empirical mode decomposition method when Intrinsic Mode Functions (IMF) are generated for complex data; second, a Hilbert transform converts the local energy and instantaneous frequency derived from the IMFs to a full energy-frequency-time distribution of the data [Huang et al. 1998]. The advantage of HHT over wavelet-based methods is claimed to be that the latter performs well on nonlinear signals with gradual inter-wave frequency modulation but poorly on signals that have intra-wave modulation (i.e., a group of signals that vary over time) while the former performs equally well on both allowing more accurate analysis.

A potential disadvantage of HHT is the necessity of its licensing for both industrial and academic applications.