METHODS FOR SEGMENTATION AND CLASSIFICATION OF SWALLOWING INSTANTS FROM THE FEEDING SOUND OF NEWBORN INFANTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ABDULLAH ONUR KOYUNCU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

JULY 2019

Approval of the thesis:

METHODS FOR SEGMENTATION AND CLASSIFICATION OF SWALLOWING INSTANTS FROM THE FEEDING SOUND OF NEWBORN INFANTS

submitted by **ABDULLAH ONUR KOYUNCU** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. İlkay Ulusoy
Head of Department, **Electrical and Electronics Engineering** _____

Prof. Dr. Tolga Çiloğlu
Supervisor, **Electrical-Electronics Eng. Dept., METU** _____

Examining Committee Members:

Prof. Dr. Çağatay Candan
Electrical and Electronics Eng. Dept., METU _____

Prof. Dr. Tolga Çiloğlu
Electrical and Electronics Eng. Dept., METU _____

Prof. Dr. A.Aydın Alatan
Electrical and Electronics Eng. Dept., METU _____

Assoc. Prof. Dr. Yeşim Serinağaoğlu Doğrusöz
Electrical and Electronics Eng. Dept., METU _____

Prof. Dr. Özgül Salor Durna
Electrical and Electronics Eng. Dept., Gazi University _____

Date: 10.07.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Abdullah Onur Koyuncu

Signature :

ABSTRACT

METHODS FOR SEGMENTATION AND CLASSIFICATION OF SWALLOWING INSTANTS FROM THE FEEDING SOUND OF NEWBORN INFANTS

Koyuncu, Abdullah Onur

M.S., Department of Electrical and Electronics Engineering

Supervisor: Prof. Dr. Tolga Çiloğlu

July 2019, 97 pages

Statistics such as swallow frequency, the average time between rhythmic swallows and the maximum number of rhythmic swallows can be related to the feeding maturity of infants. Therefore, detecting swallow segments automatically from an acoustical feeding signal can be considered as a decision support mechanism for neonatologists. This thesis includes different approaches for the analysis of infant's feeding sounds and proposes two different pattern recognition methodologies, segmentation followed by classification and classification followed by merging, for auto-segmentation and classification of swallowing instants. Data from 52 infant subjects are used, in which acoustic feeding signals are recorded with a digital stethoscope. Multiple learning algorithms such as Gaussian mixture models (GMM), support vector machines (SVM) and hidden Markov models (HMM) are used to discriminate swallowing sounds from other sound activities. A comprehensive set of feature extraction methods in time and frequency domain are investigated for the representation of captured acoustic signals. Moreover, feature selection methods are examined thoroughly to improve the repre-

sentation power of feature vectors. Experimental comparison in terms of precision, recall and F1 scores of eight different paths to segment and classify swallow instants is made. The results show that the first approach segments the swallow episodes with lower performance as the error in the segmentation also affects the classification performance negatively. On the other hand, best results are obtained in the second approach where binary and 3 class SVM classifiers are applied with purpose-specific finite state machine algorithms. In the time duration based performance evaluation, the F1 scores are obtained as almost equal to 0.70 for both methods. On the other hand, they are computed as nearly 0.81 in the event based one.

Keywords: Classification, Gaussian Mixture Model, Hidden Markov Model, Machine Learning, Newborn infants, Pattern Recognition, Support Vector Machine, Swallow Sound

ÖZ

YENİDOĞAN BEBEKLERİN BESLENME SESİ ÜZERİNDEN YUTMA ANLARINI BÖLÜTLEME VE SINIFLANDIRMA YÖNTEMLERİ

Koyuncu, Abdullah Onur

Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. Tolga Çiloğlu

Temmuz 2019 , 97 sayfa

Yutma frekansı, ritmik yutmalar arası ortalama süre, maksimum ritmik yutma sayısı gibi istatistikler bebeklerin beslenme olgunluğu ile ilişkilendirilebilir. Bu yüzden, akustik beslenme sinyali üzerinden yutma anlarının otomatik olarak tespiti, neonatologlar için bir karar destek mekanizması olarak düşünülebilir. Bu tez, yenidoğan bebeklerin beslenme seslerini analiz eden farklı yaklaşımları içermektedir ve bölütleme sonrası sınıflandırma ile sınflandırma sonrası birleştirme olmak üzere iki adet örüntü tanıma temelli yöntem sunmaktadır. Bu çalışmada, 52 bebekten dijital steteskop aracılığıyla alınan beslenme kayıtları kullanılmıştır. Yutma seslerini diğer seslerden ayırt etmek amacıyla, destek vektör makinaları, saklı Markov modelleri ve Gaussian karışım modelleri gibi bir çok öğrenme algoritmasından faydalanılmıştır. Akustik sinyalleri temsil etmesi amacıyla, zaman ve frekans uzayında kapsamlı bir öznitelik araştırması yapılmıştır. Ayrıca, öznitelik vektörlerinin temsiliyetini artırmak ve boyutunu küçültebilmek adına, öznitelik seçme prosedürleri incelenmiştir. Sekiz farklı yutma sesi bölütleme ve sınıflandırma yöntemi, kesinlik, hatırlama eğrileri göz

önünde bulundurularak deneysel olarak karşılaştırılmıştır. Elde edilen sonuçlara göre, bölütlemede yapılan hatanın sınıflandırma performansını da direkt olarak olumsuz etkilemesi sebebiyle ilk örüntü tanıma yaklaşımının daha düşük performansla çalıştığı gözlemlenmiştir. Öte yandan, 2 ve 3 sınıflı destek vektör makineleri, amaca uygun tasarlanmış birleştirme algoritmaları olan sonlu durum makineleriyle birlikte en iyi performansı sergilemişlerdir. Zamana dayalı değerlendirme yapılırken, her iki yöntem için de F1 skoru yaklaşık olarak 0.7 bulunurken, sayı temelli değerlendirmede bu değer 0.81 civarında hesaplanmıştır.

Anahtar Kelimeler: Destek Vektör Makineleri, Gaussian Karışım Modeli, Makine Öğrenmesi, Örüntü Tanıma, Sınıflandırma, Saklı Markov Modeli, Yenidoğan bebekler, Yutma sesi

To my parents and my beloved Zeynep

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Prof. Dr. Tolga Çiloğlu, for his support, guidance, instructive suggestions and comments throughout the thesis.

I am grateful to the employees of the company KuartisMED, especially to Ahmet Saraçoğlu for providing motivation, suggestions and guidance. They give me an opportunity to work on an exciting project during my graduate study.

I thank to Prof. Dr. Ayşe Nur Ecevit, Prof. Dr. Aylin Tarcan and Assoc. Prof. Deniz Anuk İnce for their instructive comments, great support and valuable helps.

I also thank to my friends, Berkay Çalışkan, Arda Deveci, Barış Karademir, Ozan Hazır, Ozan Ağma, Oğuzhan Kaya, Gökcan Bayrak, Dilem Karakaya, Ece Durmaz, Nur Akan, Taylan Yapıcı, Osman Cerlet, Ezgi Genç, Murat Özatay, Tolga Dağdelen, Gökhan Gültepe for their priceless friendship. I have always been motivated to know that they are with me.

I would like to express my deepest appreciation to my parents for their never-ending support, encouragement and unconditional love through all my life. I will do my best to be worthy of them.

My special thanks go to my wife, Zeynep Kalın Koyuncu. She was always there for me during my thesis and in my private life. I was not able to finish this work without her support, faith and love.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ALGORITHMS

ALGORITHMS

LIST OF ABBREVIATIONS

PMA	Post Menstrual Age
PNA	Post Natal Age
GA	Gestational Age
VFSS	Videofluoroscopic Swallowing Study
FEES	Fiberoptic Endoscopic Evaluation of Swallowing
DFT	Discrete Fourier Transform
PSD	Power Spectral Density
LPC	Linear Predictive Coding
MFCC	Mel Frequency Cepstral Coefficients
SVM	Support Vector Machines
RBF	Radial Basis Function
HMM	Hidden Markov Model
EM	Expectation Maximization
GMM	Gaussian Mixture Model
FSM	Finite State Machine
EMG	Electromyography
NICU	Neonatal Intensive Care Unit
VI	Variance Index
IDS	Initial Discrete Sound
BTS	Bolus Transit Sound
FDS	Final Discrete Sound

CHAPTER 1

INTRODUCTION

1.1 Motivation

Prematurity is a term for newborn babies whose births take place earlier than 37 full weeks of pregnancy. The newborns in this category are called preterm or premature infants. According to data from the World Health Organization, approximately 15 million babies, more than 10% of infants, are born preterm each year [2]. Also, the number of infants younger than five years of age who died as a result of prematurity complications is reported to be nearly 1 million. Prematurity can lead to complicated medical problems and has a negative effect on the development of infants. Swallowing disorders (pneumonia, dysphagia, etc.) and respiration-related diseases (hypoxia, respiratory standstill, etc.) can be shown as examples of short-term effects whereas systemic diseases, developmental and cognitive problems are the examples of long-term consequences. Also, preterm infants can experience crucial health problems if they are discharged from the hospital early, thus creating psychological and emotional distress for the family in addition to high medical costs. Therefore, the decision for safe discharge from the hospital is of great importance.

Minimizing the adverse effects of prematurity problems highly depends on the correct assessment of the oral feeding skills of the babies. Since the majority of preterm infants suffer from oral feeding difficulties, a lot of research has been done in order to extract the causes of nutritional problems and better assist infants accordingly.

To better understand whether or not successful and safe oral feeding skill of the baby is achieved and to discharge babies from the hospital at the right time, the doctors strongly need decision support mechanisms. In this case, the ultimate goal is to ana-

lyze and evaluate the feeding process correctly. In other words, improving the reliability of diagnoses on oral feeding difficulties and decisions for required interventions to increase success in oral nutrition are the priorities of neonatal doctors. The feeding process can be divided into three fundamental phases: suction, swallowing and respiration [3]. Although each stage has its own responsibility in the process of taking milk from the oral cavity and delivering to the stomach safely, the synchronization of them is also of great importance. Because the lack of timing may cause the accidental deposit of food to the lungs and respiratory standstill. Hence, developmental, cognitive and neurological problems may occur.

In order to estimate swallowing actions quantitatively and analyze the characteristics of that, a lot of methods are proposed in the previous research studies for both adults and infants. Although videofluoroscopic swallowing study (VFSS) [4] and fiberoptic endoscopic evaluation of swallowing (FEES) are successful in capturing the swallowing moments and helping doctors interpreting nutritional maturity, they are invasive diagnostic procedures requiring an invasion of the body cavity or disruption of regular body functions. On the other hand, the swallowing function is tried to be evaluated through probes placed in human skin such as electromyography (EMG) electrodes, accelerometers, contact microphones, stethoscopes and piezo-resistive pressure sensors. Although these methods are non-invasive, they are not as strong as mentioned invasive techniques in providing evidence-based support for neonatal doctors.

1.2 The purpose of this study

In this study, acoustic feeding signals are taken from [5] and they are recorded from infant subjects by a digital stethoscope. Swallowing action is monitored, analyzed and evaluated. Different acoustic swallow patterns together with other types of non-swallow sounds (click, smacking, respiration, etc.) produced by infants are analyzed. As a result of these examinations, two approaches containing digital signal processing techniques and conventional machine learning algorithms are proposed to detect swallow episodes automatically from the acoustic feeding signal. The proposed method can be considered as a solution for an acoustic semantic segmentation problem in which only swallow moments are segmented.

Objectives of both approaches are the same, yet they differ in the pre-processing and post-processing stages of the whole process. Besides, these solutions contain modules in which various techniques and parameters are utilized to optimize detection performance. Moreover, a large-scale experimental comparison of machine learning-based solutions for the problem of automatic swallow event detection is made to evaluate classification performances on swallow sounds.

Swallow maturity can be considered as one of the most significant indicators of oral feeding readiness for newborn babies. Previous research studies have shown that the swallowing frequency is positively correlated with feeding performance. Similarly, experimental findings pointed out that postmenstrual age (PMA) and the average time between rhythmic swallows have a negative correlation [5]. In addition, an increase in the maximum number of rhythmic swallows refers to the development of feeding skills of infants. Thanks to the proposed method, statistical data obtained from detected swallow segments can be related to oral feeding maturity of the infants, hence assist neonatal physicians to decide it.

1.3 Thesis Organization

In Chapter 2, a literature review about oral feeding maturity, swallow physiology and multiple techniques for swallowing behavior analysis and evaluation of both adults and infants are provided in detail. The studies expressing the impact of different parameters on oral feeding readiness of the infants are included. Both invasive and non-invasive assessment techniques of swallowing function for all target groups are expressed.

In Chapter 3, information about the data set is given. The preparation of training and test data sets and labeling rules are explained. Different types of swallow patterns and the spectral characteristics of swallow and non-swallow sound segments are shown.

Chapter 4 presents the pattern recognition methodologies for the detection of swallowing events. Based on this methodology, the design procedure of the two approaches, segmentation followed by classification and classification followed by merging are described. Segmentation, feature extraction, classification and merging meth-

ods are expressed in detail. Pseudo codes of train and test procedures are given for each machine learning algorithm.

Chapter 5 presents the implementation details together with the results of segmentation, feature extraction and the classification modules. The experimental procedure, including all the modules, is given. Selection criteria for appropriate features are described. Performance calculation and parameter optimization methods are explained for segmentation, classification and merging parts separately. Experimental comparison in terms of precision, recall, and F1 scores of eight different paths to detect swallow episodes are tabulated.

Chapter 6 summarizes the work done in this study and highlights the crucial points of the thesis. Concluding remarks are presented and future work suggestions are given in this chapter.

CHAPTER 2

LITERATURE REVIEW

In this chapter, firstly, the studies related the factors affecting the oral feeding maturation and information on the physiology of swallowing function are told. Secondly, both invasive and non-invasive assessment methods of oral feeding readiness are discussed.

2.1 Swallow Physiology & Oral Feeding Maturity

Swallowing can be considered as a procedure in which several muscles and nerves work together and it has three major phases, namely, oral, pharyngeal and esophageal. Receiving food from the mouth, chewing and softening the food with saliva to make swallow easier happen in the oral phase. In addition, the prepared food material is transferred to the behind the oral cavity with the help of the tongue. This stage ends up with the triggering of the pharyngeal phase. In the pharyngeal phase, respiratory functions are stopped to prevent the escape of food material into the trachea or airways. The esophageal phase is initiated with the entry of bolus and is the period which ends up when the liquid or food reaches to stomach [6]. The swallow procedure is illustrated in Figure 2.1.

Lau et al. [8] characterized the swallowing function, as a complex behavior requiring proper coordination of muscles from mouth, palate, pharynx, larynx, and esophagus. One of the main highlights of this study is that the swallowing function activation is not directly dependent on sucking and it is emphasized that coordination of swallowing and sucking functions is of great importance for oral feeding readiness.

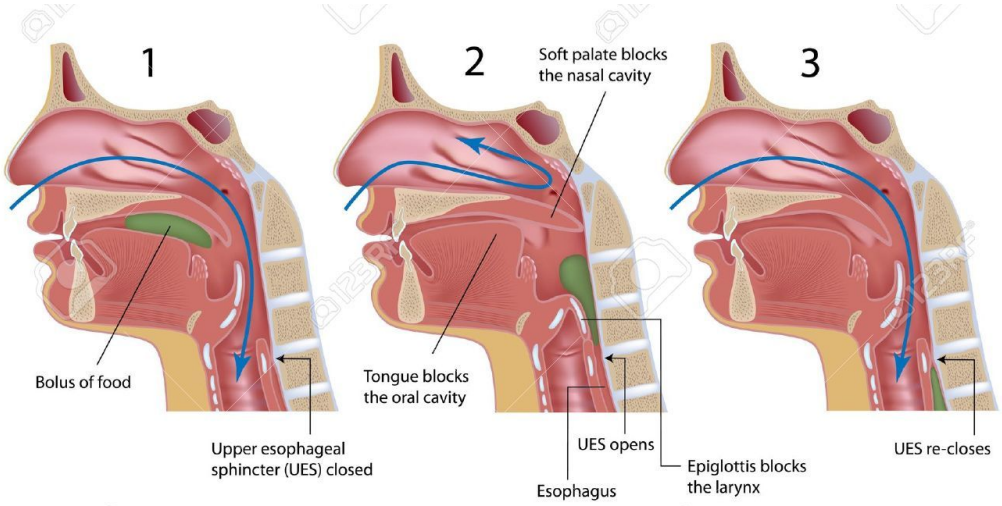


Figure 2.1: Swallow phases [7]

.

Bu et al. [9] carried out a study to reveal the relationship between optimal oral feeding readiness and the synchronized suck-swallow-breathe coordination ratio. When sucking-swallow and swallowing-respiration pairs were examined separately, it was observed that the coordination ratio converges to 1:1:1 sequence as GA (gestational age) increases. Although there were no definite findings in the literature or their research, they stated that there might be a relationship between the coordination of tongue movements and the converging to 1:1:1 sequence.

Lau et al. [3] claimed that larger swallow frequency, increase in the bolus size, advanced swallowing and respiration coordination enhance the milk transfer ratio. Thus, they aimed to examine the coordination of suck-swallow and swallow-breathe pairs separately. As a result of this study, the following results were found:

- The milk transfer ratio raises over time and has correlations with both average bolus size and swallow frequency.
- Average bolus size is correlated with suction amplitude.
- Average bolus size is not correlated with swallow rate.
- Swallow frequency is correlated with the frequency of sucking.

In addition, the swallows that occurred at different phases of the respiration cycle

were observed. Experimental results depicted that, occurrence time of swallow is related to the readiness of oral feeding.

Gewolb et al. [10] investigated the relation of PMA and post-natal age (PNA) with defined oral feeding readiness parameters such as stability of sucking rhythm (defined as a function of sucking interval mean) and swallow rhythm. The authors pointed out that, mentioned feeding-related parameters are correlated with PMA significantly. However, none of the measurements is correlated with PNA. In addition, they argued that since PMA is a better indicator of feeding patterns than PNA, feeding patterns can be considered as not learned but innate behaviors.

Amaziu et al. [11] assumed that significant difficulties in oral feeding arise from the development of related muscles at different times in premature infants. By monitoring the necessary parameters related to feeding ability, they deduced that sucking, swallowing, respiration and coordination of them are matured at different rates and times. Furthermore, in order to determine the maturation level of infants, their gestation was found to be more informative than PMA.

Lau et al. [12] conducted a study to determine the effect of defined feeding parameters on oral feeding readiness level. In this study, PRO (what percent of predetermined milk volume is consumed in the first 5 minutes), RT (milk flow rate during entire nutrition period), GA, SOF-IOF (duration between the start to independent oral feeding) were chosen as the major components to determine oral feeding skills of infants. And OFS is the combination of PRO and RT. Results of this study given below concluded that OFS could be used as a new objective representation of infants skill and durability.

- OFS levels are correlated with GA, PRO and SOF-IOF
- SOF-IOF is associated with GA and OFS.
- RT is only correlated with OFS.

Lau et al. [13] stated that oral feeding maturity is better defined with the coordination of suck-swallowing process-breathing than suck-swallow-breathe as swallowing process includes not only pharyngeal phase of the swallowing but also oral and

esophageal phases. Hence, any levels of a nutritive sucking pathway may cause an inefficient or unsafe feeding procedure since different components and muscles within corresponding levels mature at different times. In addition, it was argued that such occurrences can be considered as a reason why infants with the same PMA and GA differ in terms of readiness to oral feeding.

2.2 Assesment of Swallow Function

In this section, the background studies related to the instrumental evaluation techniques of swallow function are described. In addition, both invasive and non-invasive examination methods to guide for diagnostic procedures and therapeutic decisions are proposed.

2.2.1 Invasive Methods

2.2.1.1 Videofluoroscopic Swallowing Study

The videofluoroscopic swallowing study (VFSS) or modified barium swallowing examination (MBS) is a method for evaluating the swallowing physiology using the form of a real-time x-ray. To evaluate swallowing ability, a bolus form including the range of food or liquid consistencies and barium is given to patient [14]. In Figure 2.2, a young male's normal swallowing events are illustrated fluoroscopically.

The VFSS is a tool for management, characterization and evaluation of swallow function. Indeed, it is known as the gold standard for diagnosing the pharyngeal swallow impairment and aspiration to the lower airways [15, 16].

Using the clinical features of the VFSS results as an input, fewer studies are conducted to detect swallowing and respiratory abnormalities in infants compared to adults. Newman et al. [17] aimed to figure out how the occurrence frequency of the laryngeal penetration and aspiration in infants having dysphagia can relate to VFSS findings. The study indicated the abnormalities such as laryngeal penetration and aspiration on the infants suspected of having swallowing difficulties. However, they

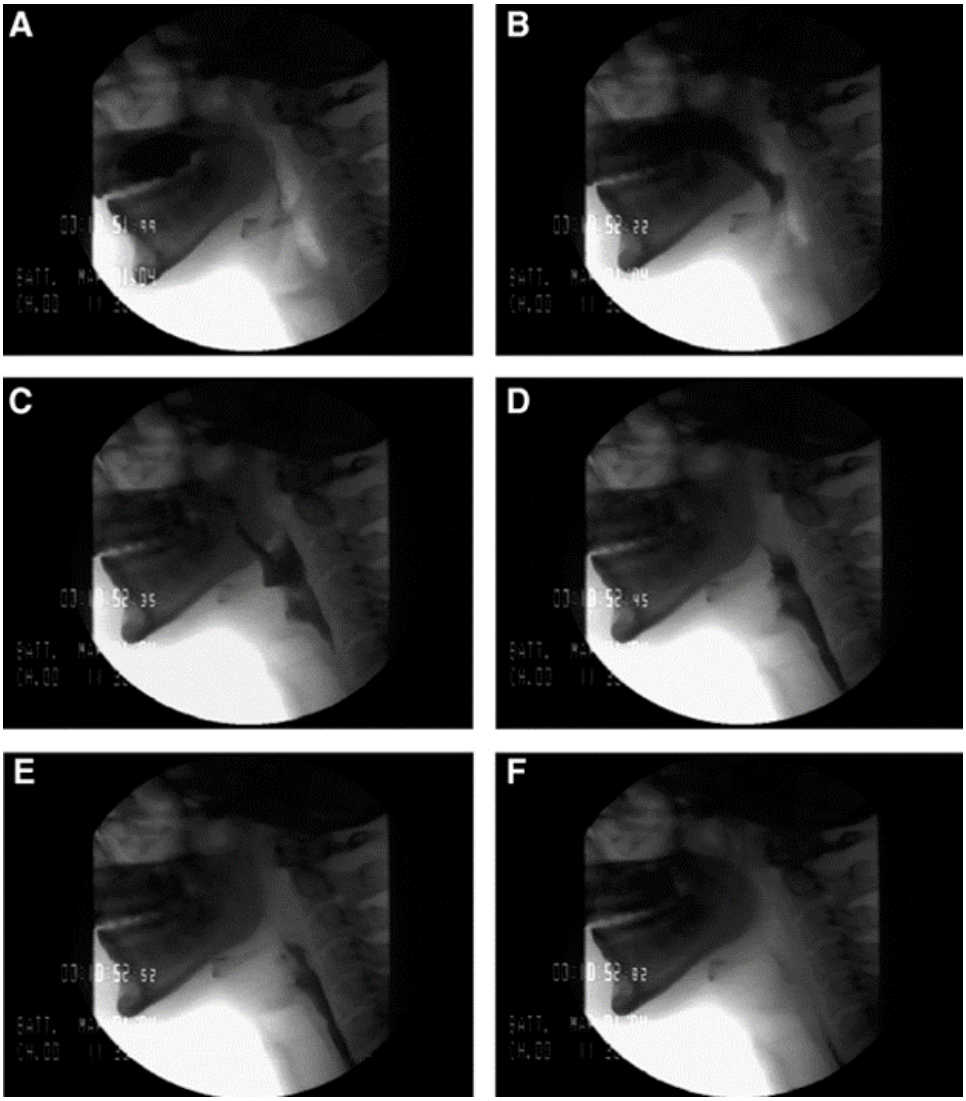


Figure 2.2: VFSS images of an adult's swallow procedure in which the bolus is transferred from mouth to the stomach [14]

.

have stated that multiple swallows must be performed to make a radiologic assessment.

Kim et al. [18] investigated the prediction performance of the VFSS study on the infants of two groups called aspiration and pneumonia group. The first group who had no pneumonia and having aspiration symptoms showed abnormalities only in pharyngeal phases, whereas the VFSS findings revealed anomalies in both oral and pharyngeal phases in the second category who had pneumonia.

In order to evaluate the differences between full-term and preterm infants having dysphagia, Uhm et al. [19] utilized VFSS as a referral. Although there was no significant difference in the VFSS findings of preterm and full-term infants, the decrease in the sucking speed was more frequently in the preterm infants.

2.2.1.2 Fiberoptic Endoscopic Examination of Swallowing Study

The fiber-optic endoscopic examination of swallowing study (FEES) is another tool for evaluation of the pharyngeal stage of swallowing action in patients with dysphagia [20]. In this procedure, a fiber-optic scope extending from the patient's nose to the pharynx is inserted. In the next step, a physician evaluates the swallowing process by observing the scope when the patient is fed.

In Figure 2.3, two different endoscopic views rendered from an old version camera and high-resolution system are shown. Previously, although this procedure was approached with suspicion, it later experienced a dramatic change and began to be accepted as a primary method for suspected patients with dysphagia [21].

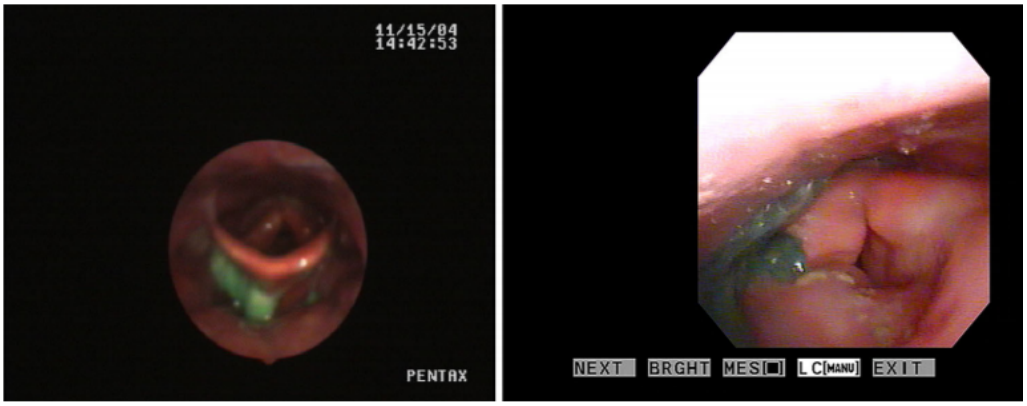


Figure 2.3: Fiber- optic scope views captured by an old version camera (left), high resolution camera (right) [21]

In order to analyze and evaluate the swallowing function in the breastfeeding infants, Willete et al. [22] used FEES as a safe and effective instrument. They also stated that other instrumentation tools like VFSS are not able to assess the dysphagia during breastfeeding in the target group of the study including 23 infants with the average age of 14 weeks and two of them are premature.

In their study, Reynolds et al. [23] provided an overview of current challenges in assessing aspiration for adult, pediatric, infant and neonatal populations and presents a multidisciplinary FEES program for feeding and breastfeeding in the Neonatal Intensive Care Unit (NICU). The NICU feeding team constructed a NICU FEES program consisting of 5 main components: equipment, education, competency, protocol and procedure. They concluded that FEES could be considered as a safe and effective alternative to the VFSS. However, the current research is not enough to compare the validity and efficacy of the VFSS and FEES.

2.2.2 Non-Invasive Methods

2.2.2.1 Ultrasound Swallow Study

Ultrasound imaging is a non-invasive tool to visualize inside of the body sound waves with high frequencies. This technology is utilized to examine many of the internal organs such as heart, liver, kidneys, uterus, etc., as well as to observe muscles and elements involved in swallowing action.

To explain the events, that occurred during swallowing action, Weber et al. [24] benefited from ultrasound technology. An ultrasound scanner was used with a video recorder to assist clinicians in evaluating the sucking and swallowing dynamics of the newborn infants and the coordination of sucking, swallowing and respiration. By observing the muscle movements during swallowing, they depicted that milk availability significantly affects the ratio between suck and swallow. In addition, increasing PNA also increases the frequency of swallowing. Based on the observations, they suggested that the feeding mechanism can be appropriately studied and investigated by ultrasound technology and the scope of the study can be extended in the future.

Using similar instruments, Bulock et al. [9] found out that neuromuscular maturity implies maturity in the nutrition mechanism. Moreover, the excess of the gestation period affects the feeding more effectively than PNA does.

Geddes et al. [25] aimed to visualize swallowing function via ultrasound approach and made a comparison between respiratory inductive plethysmography (RIP) and

this approach. In this study, the results showed these methods are highly correlated. Furthermore, the pharyngeal phase of the swallowing can be observed, independent of which the respiration phase is. In Figure 2.4, the movement of the milk bolus during breastfeeding can be seen.

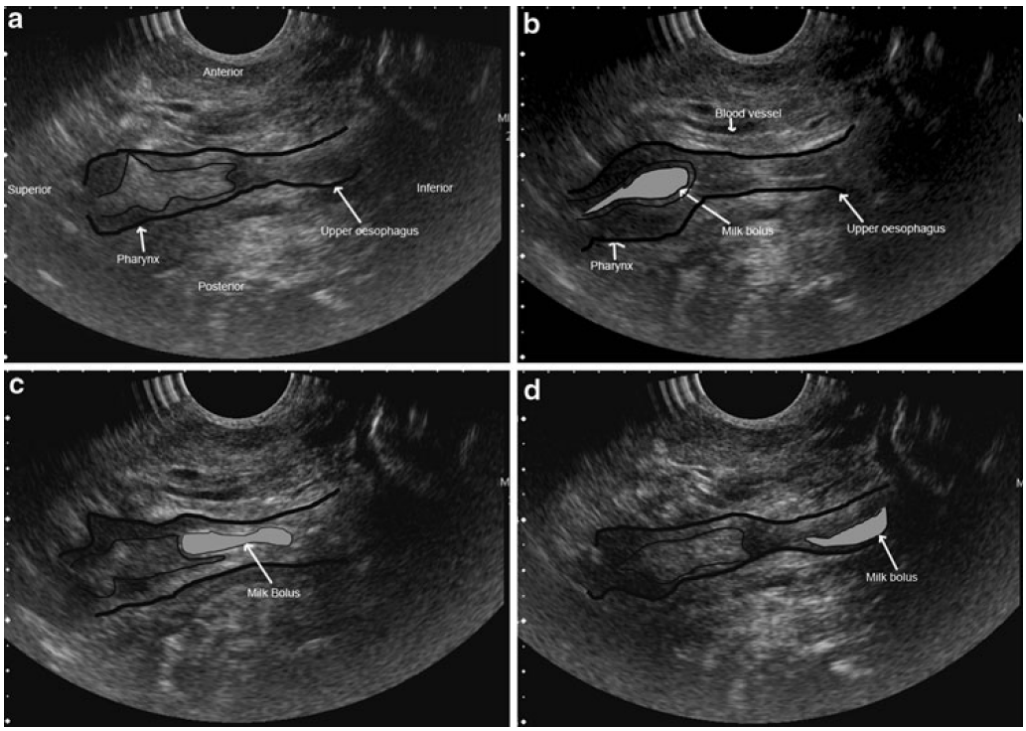


Figure 2.4: View of the milk bolus movement during breast-feeding from an ultrasound scanner [25].

Sonies et al. investigated many components of the ultrasound technology in terms of techniques and procedures for different target groups. Additively, their study also discussed the disadvantages and limitations of the technology, although it has many advantages such as being non-invasive, safe, portable and is very useful in swallowing applications such as diagnosing swallow impairments, determining swallow duration, evaluating infants during feeding, etc. The most important one is that bones involving during the swallow cannot be visualized since the high-frequency sound waves are not able to pass through. Therefore, the scanning region is limited due to the presence of bones. Hence, this technique may lack to detect aspiration moments.

2.2.2.2 Other methods

In this part, studies focusing on pressure and acoustic-based measurement techniques for the identification of both adult and infant swallowing sounds and acoustical modeling of swallow mechanism are discussed.

Takahashi et al. investigated the swallowing sound detection subject on adults from three different aspects and figured out the following results [26].

1. **Type of Acoustic Sensor:** The accelerometer was selected as a proper transducer since the amplitude values of the frequency spectrum in a wide range are large enough to work properly with a low attenuation level.

2. **Type of Adhesive:** Double-sided paper tape was used to attach accelerometer to the skin.

3. **Optimal Location:** Swallowing sounds were acquired at 24 different regions of the neck to select the optimum location. Performance evaluation was done based on observing the signal-to-noise ratio. As a result, the lower region of the cricoid cartilage and the upper side of the trachea lateral boundary were found to be the best location.

Cichero et al. [27] found inconsistent results concerning the paper written by Takahashi et al. [26] in terms of the acoustic transducer to use for detection of swallowing by revisiting the methodology. Similarly, they used the signal-to-noise ratio as a performance evaluation parameter. Based on the findings of their study, they argued that the electret microphone should be preferred rather than the accelerometer for recording swallow sounds. However, they did not object to findings of optimal location.

Reynolds et al. [28] defined the acoustic signals due to pharyngeal movements of infants occurred at the beginning of the swallow action as initial discrete sound (IDS). They calculated the “variance index” (VI) parameter of each infant from the captured waveform with the help of digital signal processing technology. Using this term, they found that regularity of swallow sounds of the preterm infant goes up as PMA increases. Furthermore, they suggested that the methodology used in this paper could

associate with the feeding maturity. Another important issue is the determination of the acoustic device and the location where it should be placed. In Figure 2.5, the picture of the accelerometer and representation of attachment were illustrated.

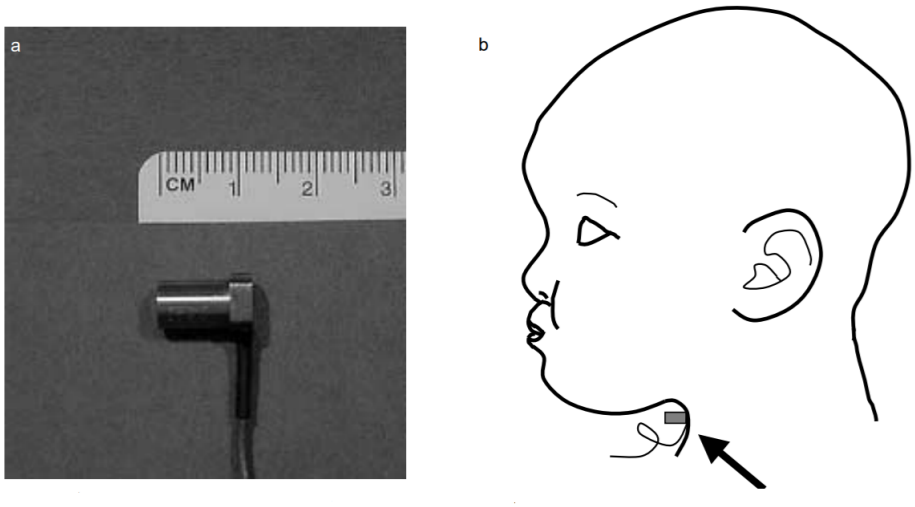


Figure 2.5: Illustration of accelerometer placement to the neck of an infant

Amft et al. [29] studied to detect and classify swallowing activities of adults by processing the acoustical signal recorded via electret condenser microphones placed inferior mid-line from the cricoid cartilage. For the same purpose, they also observed the muscle movements quantitatively with EMG sensors attached to the throat. The position of both sensors is depicted in Figure 2.6. Two different methods, feature similarity and the signal intensity, were compared to extract swallow instants. The latter method using the EMG signals showed better performance than the sound signal obtained from the microphone. However, when the results of the two sensor data were fused, the accuracy was increased.

Reynolds et al. [30] extended their study about evaluation of IDS stability [28] by including adults as well. As a result, the VIs of the adults and premature babies older than 36 weeks PMA were not found to be different. On the other hand, greater VI values were obtained when the PMA was decreased below 36 weeks. Besides, two sensors, accelerometer and microphone, were compared in real-time for the first time and deduced that both can be used to analyze swallowing-related sounds and that in some special applications, one can be preferred to the other.

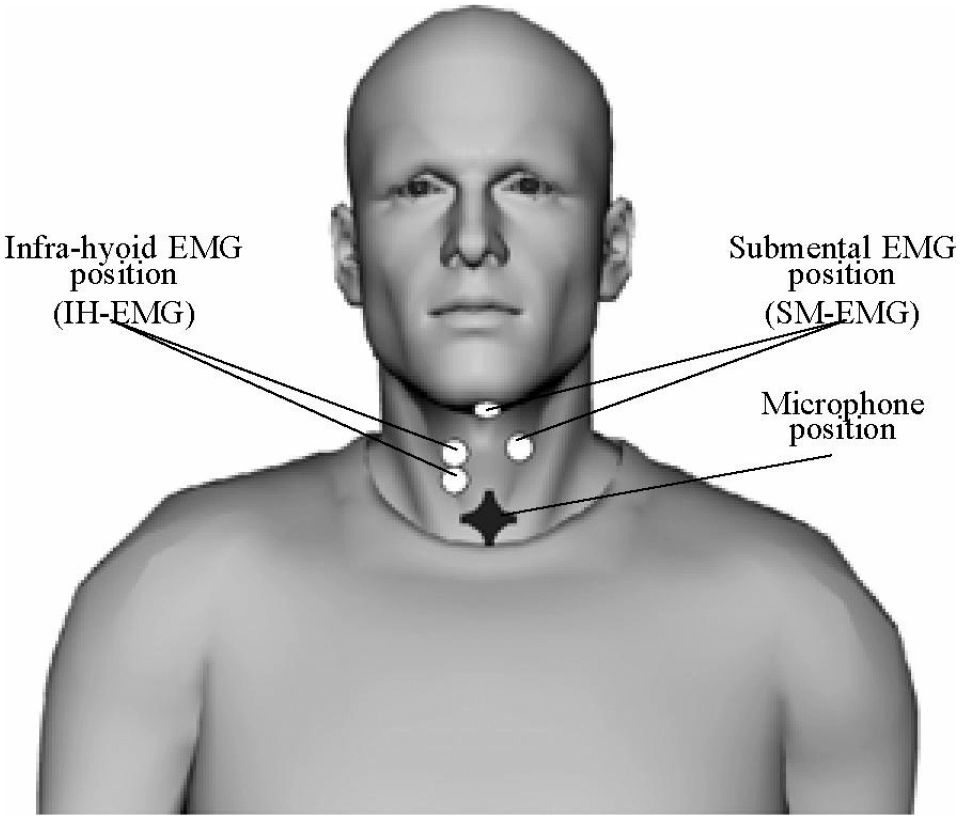


Figure 2.6: Illustration of electret condenser microphone and EMG sensors attachment on an adult's neck

At the University of Manitoba, Zahra Moussavi and her colleagues conducted several studies to acoustically analyze and evaluate the swallowing mechanism for children and adults. And, the majority of them involve signal processing and machine learning techniques. Swallow sounds were acoustically captured and processed to show the differences between normal (having no swallowing difficulties) and people with dysphagia [31]. In the beginning, swallow sounds were extracted with the help of distance-based features. Then, segmented swallow activities were classified, employing a discriminant algorithm. In another study, they assumed that the wavelet transform coefficients should be high for each scale due to the non-stationary nature of the swallowing sounds [32]. For that reason, the captured acoustic signal was exposed to a wavelet-based filter. The identification of swallow action from respiratory events was carried out in an automated way. The results were validated by visual inspection of the filtered signal and airflow measurements. In Figure 2.7, the original (from the acoustic transducer) and filtered signal were shown. As it is seen, the breath sig-

nal has smaller components in higher wavelet scales while swallow sound remains alive on that scale. After that, another solution was found to automatically distinguish swallowing and breathing sounds with the help of an HMM-based classifier. Thanks to HMM-based classifier together with the recurrence plot features which is reconstructed via state space trajectories of the swallow and respiration sound signal, detection performance developed to a level better than the former studies [33]. To relate sounds of respiration events occurring right after the swallow actions with dysphagia, a novel method was identified in another study [34]. First of all, the time domain signal of the breathing action together with the first and second derivatives were plotted in a 3-D space. Secondly, the scattered 3-D data fitted an ellipsoid to discriminate breathing events after the disordered swallow moments. Finally, the mean of data outside the ellipsoid was used as a feature of corresponding respiration interval to be given as an input the SVM (Support Vector Machine) classifier. The accuracy was found to be 86% in a target group consisting of 50 adults with dysphagia.

Youmans et al. [35] characterized the acoustic signals recorded via cervical auscultation. The study was carried out by listening to the sounds of swallowing with the help of a stethoscope placed over the patient's neck. Swallow sounds of healthy adults with a wide age range were recorded during the ingestion of boluses with different viscosity and volume. The results of this study showed that the change in the swallowing characteristics is more related to the viscosity rather than volume. Also, the increase in the duration and decrease in the intensity of the captured signal was found to be correlated with increasing age.

In order to monitor three main elements of the feeding mechanism (sucking, respiration and swallowing) to observe coordination between, Chen et al. [1] developed a wireless system for preterm and term infants. A pressure sensor was placed on a feeding bottle to digitize the sucking pressure. In addition, breathing signal was obtained via two EMG neonatal electrodes while swallow data acquired with the help of a mini microphone. Each probe was connected to a wireless acquisition module and passed through simple filters after digitization. The detection algorithm of the sucking, respiration and swallow activities were built in a back-end system that receives the output of the wireless module via Bluetooth. This purpose-designed system is depicted in Figure 2.8. To determine corresponding activities, fractal dimension, which

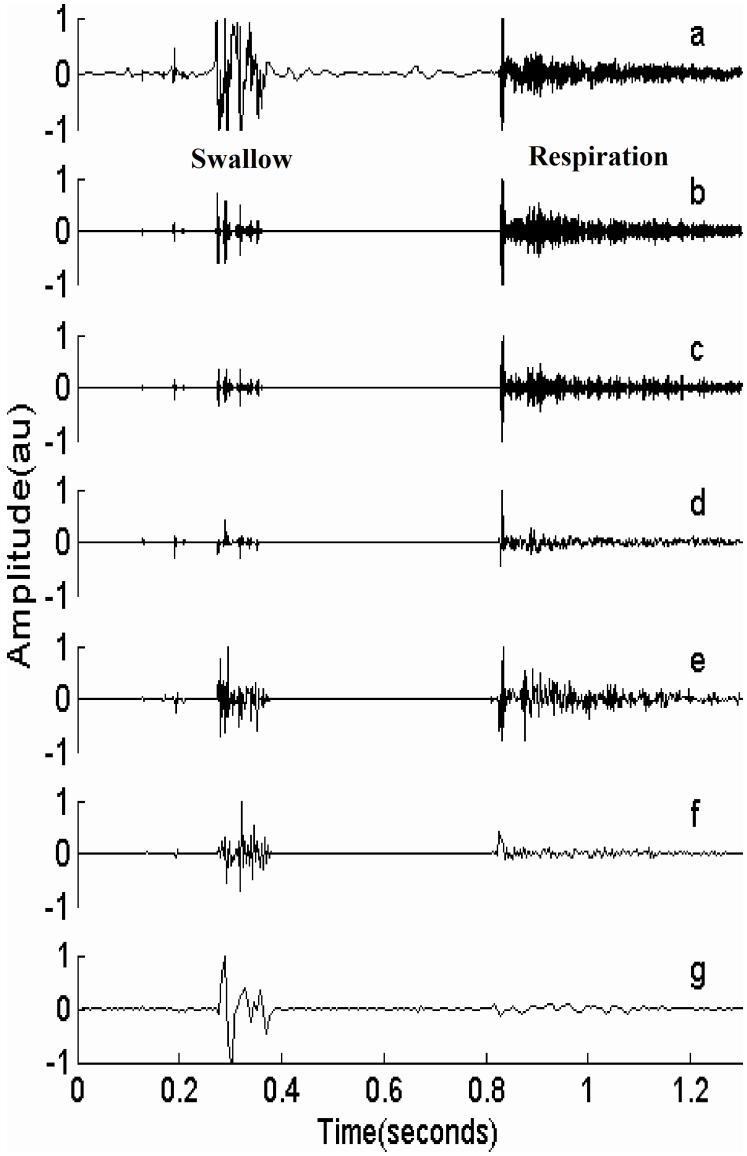


Figure 2.7: Decomposition of the tracheal sound signal including both swallowing and respiration events [32]

is related to the complexity of the signal, was extracted for EMG and microphone signals. After that, a simple peak detection algorithm (first-derivative based) for the fractal dimensions and the pressure signal related to the sucking event was applied to those signals. The performance parameters, sensitivity and the positive predictive value (PPV), were calculated over 85%. Then, this study was modified [36] by using an optical probe instead of EMG probes to monitor abdominal movements during breathing. Together with that development, performance for the respiratory event detection slightly increased.

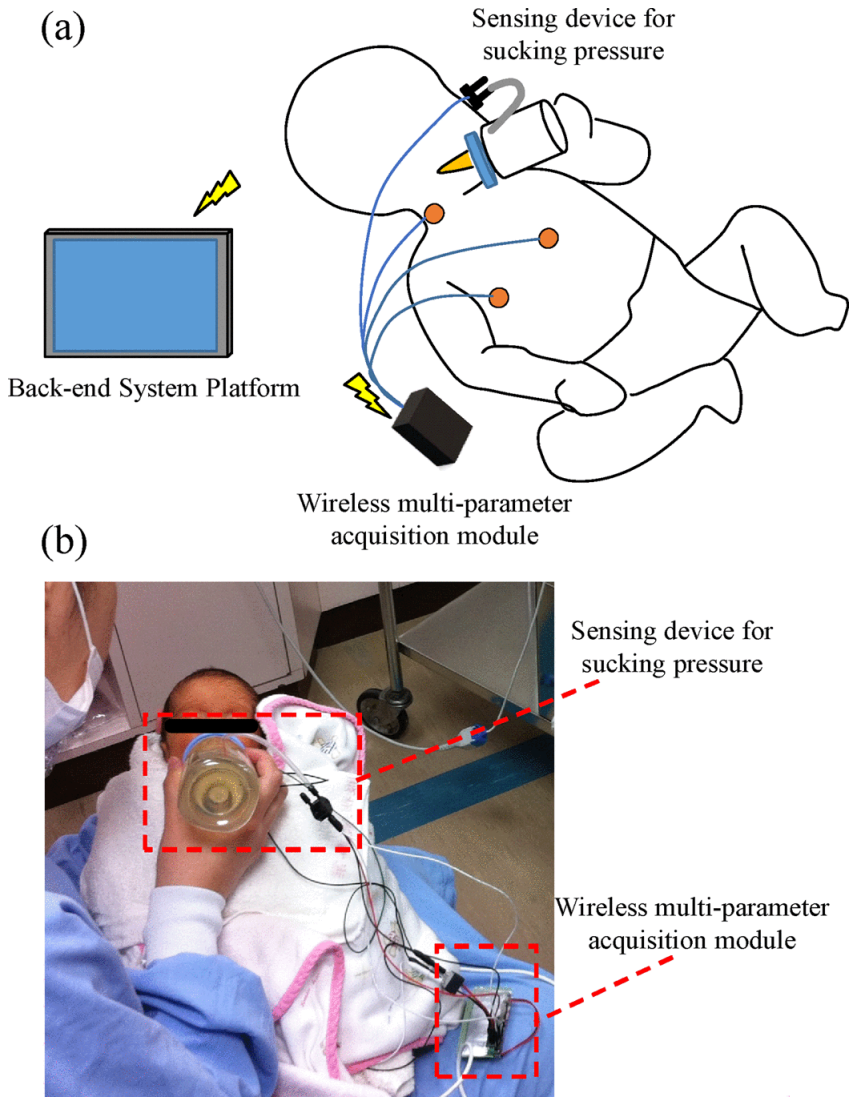


Figure 2.8: A wireless system to assess the characteristics of sucking, swallow and respiration for infants [1]

Ince et al. [5] investigated the feeding maturation of infants by evaluating the swallow sound signals captured from a digital stethoscope. Variables such as total number swallows, the maximum number of rhythmic swallows, milk volume, etc. were generated from two-minute audio recordings to observe correlation with maturation. In this study, 52 preterm and 42 full-term infants were auscultated. During auscultation, swallow activities were labeled by a specialist in real time and synchronization mismatches were corrected immediately after the feeding session was over. As a result, a positive correlation was found between PMA and both the maximum number of rhythmic swallows and volume of milk ingested. Moreover, the increase in the

PMA minimizes differences between term and preterm group in terms of maturation as many studies related to this topic stated.

2.2.3 Comparison of methods

In the previous sections, several assessments and analysis methods of feeding mechanism components given in the literature are categorized, whether they are invasive or not. In this section, the advantages and disadvantages of the mentioned methods will be discussed and a comparison will be made accordingly.

From the background study, it is inferred that the invasive methods, VFSS and FEES, are accepted as the gold standard for evaluation of the swallow function, diagnosis of swallowing impairments and analysis for the clinicians. The comparison of non-invasive methods with invasive ones to test and validate their performance supports this deduction. On the other hand, researchers still seek more straightforward and practical solutions even though the aforementioned invasive methods are seen as powerful assessment tools. Table 2.1 indicates the advantages and disadvantages of invasive and non-invasive techniques.

Table 2.1: Comparison of Feeding Maturity Evaluation Techniques

Method	Procedure	Mobile	Painful	Frequency of Usage	Evaluation	Training	Other
VFSS	Invasive	No	Yes	Just for diagnosis	Complex	Moderate Level	Radiation
FEES	Invasive	No	Yes	Just for diagnosis	Complex	Expert Level	
Ultrasound	Non-invasive	Yes	No	Anytime	Moderate	Moderate Level	
Cervical Auscultation	Non-invasive	Yes	No	Anytime	Simple	Expert Level	

Although there are many studies regarding cervical auscultation to analyze the characteristics of swallow sound of adults, the research for infant subjects remains limited. However, due to the disadvantages of invasive methods, assessment of swallow function by acoustical means is of great importance, especially for newborn babies. The authors of [37, 10, 5], investigated the swallow-related sounds of newborn infants and some tried to associate to the feeding maturity of infants with different digital signal processing techniques.

Automatic swallow segmentation algorithms of adult subjects are available in previous studies [38, 39]. Also, in [36], swallow events of infants subjects were detected yet the authors did not propose any method regarding the determination of boundary limits. In this study, swallow events of infants are segmented automatically and both the onset and end limits of swallow actions are found to extract statistical results for evaluation of oral feeding readiness by using acoustic feeding recordings taken from [5].

CHAPTER 3

DATA PREPARATION & SWALLOW SOUND ANALYSIS

3.1 Formation of Training and Test Sets

Since the objective of this study is to detect swallow sound segments of infant subjects automatically with the help of machine learning algorithms, a data driven-system is required.

For that reason, acoustic feeding signals which are sampled at 44:1 kHz and recorded via a digital stethoscope (ds32a, thinklabs) in [20] were partially included the data [5] set of this study. The length of each feeding signal is two minutes. Each feeding recording was ca ptured in a quiet environment when the stethoscope was held to the hyoid region of infant subjects. This study aims to create a swallow episode detection system. For that reason, acoustic swallow patterns, swallow frequencies, resting intervals, of infants from 27 weeks 36 weeks, were visually inspected and analyzed. As a result, it is decided that infants older than 36 weeks were appropriate for the purpose of this study (52 recordings in total).

In addition to feeding recordings, text files including the beginnings of the swallow events for each baby are available. These text files were generated and written during the feeding session by a software. Each time a swallow event occurred, a specialist doctor clicked the mouse to specify swallow time. Then, audio and text

les were analyzed subsequently to correct possible time synchronization mismatches. Finally, labeling was done via Praat (a speech analysis tool) [40].

In addition to swallowing events labeled as 'y', several non-swallow sound activities were observed and they were labeled as 'n' (non-swallow) as indicated in Figure

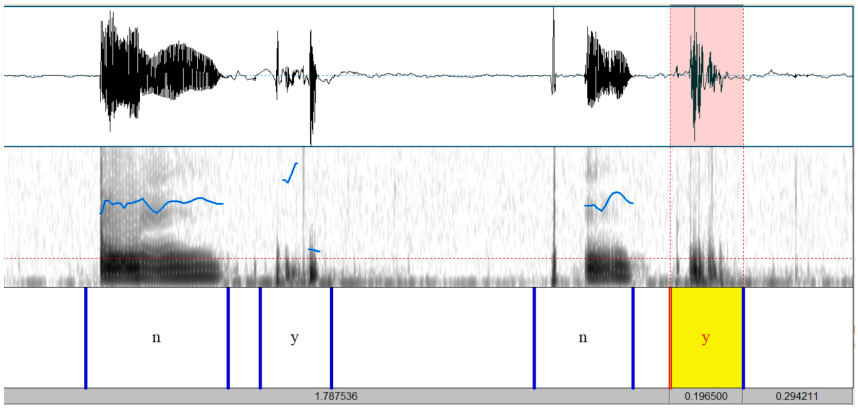


Figure 3.1: Praat view of two episodes for both "y" and "n" classes

3.1. Moreover, if there exists a sound episode that cannot be perceived as any of the classes, it has been marked as ambiguous ('a'). In the end, the number of intervals labeled as 'y' is 1003, whereas it is 837 for non-swallow. Then, randomly selected 17 feeding signals were included in the test data set and the remaining 35 audio files in the train data set. In Table 3.1, the number of intervals for both swallow and non-swallow classes for the test and train data set are depicted.

Table 3.1: Number of test and train intervals for both classes

	Interval Number			**File Number**		
	Whole	**Training**	**Test**	**Whole**	**Training**	**Test**
Swallow ('y')	1003	642	361	-	-	-
Non-swallow ('n')	837	579	258	-	-	-
Total	1840	1221	619	52	35	17

3.2 Analysis of Swallow Sounds

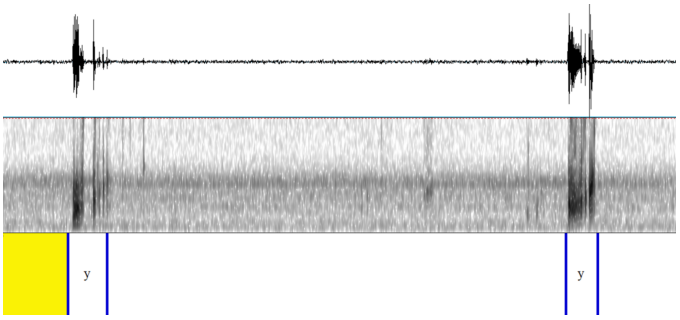
3.2.1 Swallow Patterns

Even though there are similarities, it is quite difficult to say that the swallowing sounds of this study belong to a particular pattern. Three terms related to the swal-

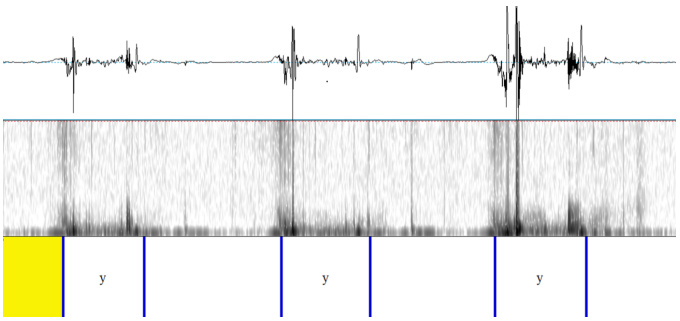
lowing process were defined in the previous studies [37, 28]: initial discrete sound (IDS), bolus transit sound (BTS) and final discrete sound (FDS). In this study, it was observed that there are similar sequences, but the presence of FDS and IDS was not guaranteed. Besides, inspiration and expiration sounds may appear before or after the bolus transmission event. However, BTS was considered to be permanent that is why only it was labeled as a swallow interval. In Figure 3.2, four different swallow sound patterns are indicated. The meaning of labels can be seen in the following list.

- **y:** swallow sound label (swallow class)
- **fds:** final discrete sound label (non-swallow class)
- **rsp:** respiration sound label (non-swallow class)
- **n:** vowel, pleasure or crying sounds label (non-swallow class)

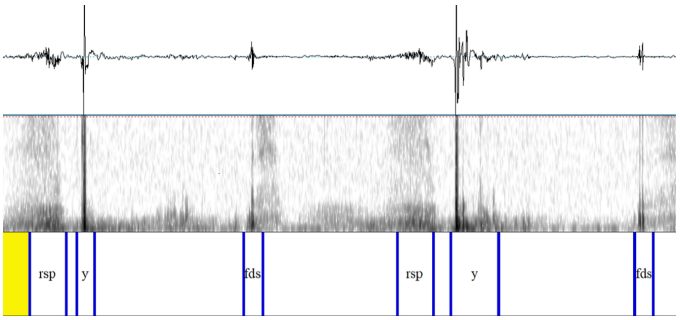
As it can be inferred from Figure 3.2, no unique swallowing process can be identified for this study. Although the existence of inspiration, expiration, FDS and IDS sound give an idea of the presence of swallow events, they were treated as non-swallow as the numbers of intervals for given interval types are limited to apply machine learning techniques.



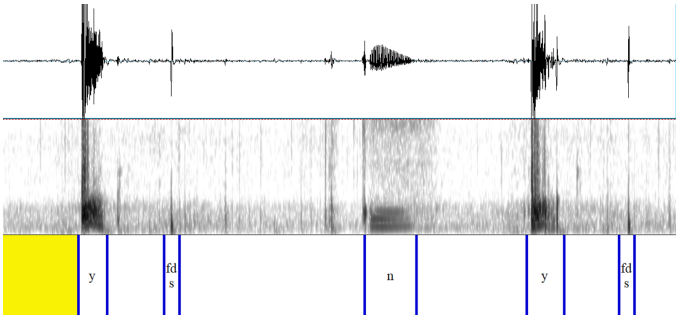
(a) Two swallow events in a compact form, no initial or final sounds



(b) Three swallow events including two stages, short silence period between stages



(c) Two swallow events, the "rsp-y-fds" pattern



(d) Two swallow events, the "y-fds" pattern

Figure 3.2: Four different swallow patterns

3.2.2 Frequency Analysis

After all the intervals were labeled, the frequency content of swallowing episodes and different types of non-swallowing sounds (pleasure sounds, cry for assistance) were analyzed. First of all, the time-frequency characteristics of signals from both classes were observed. In Figure 3.3, magnitude squared of the short time Fourier transform (spectrogram) of a sample signal including two swallow and two non-swallow intervals is depicted. As it is seen, three frequency bands seem to be dominant over the spectrum since non-swallow sound episodes produced by an infant is tonal-like or pitchy. On the other hand, swallow example has a more flat frequency response.

In addition, power spectral density (PSD) analysis was performed to examine the frequency characteristic of swallowing sounds. PSD function returns variations of power in terms of frequency and it's periodogram estimate formula for a discrete waveform, x, of sample length N and sampling frequency f_s, is given as

$$\hat{P}(f) = \frac{1}{Nf_s} \left| \sum_{n=0}^{N-1} x_n e^{-j2\pi fn} \right|^2 . \tag{31}$$

Periodogram PSD estimates of swallow and non-swallow activities are plotted in Figure 3.4. Since vowel sounds of newborns are the most common non-swallow types, it is important to compare two classes from the frequency representations as in this case. Previous studies state that swallow sounds are more complex and have higher frequency contents compared to respiration sounds. The (c) part of Figure 3.2 can also be interpreted similarly. Hence, a powerful classifier will be able to distinguish two classes with the help of frequency domain features.

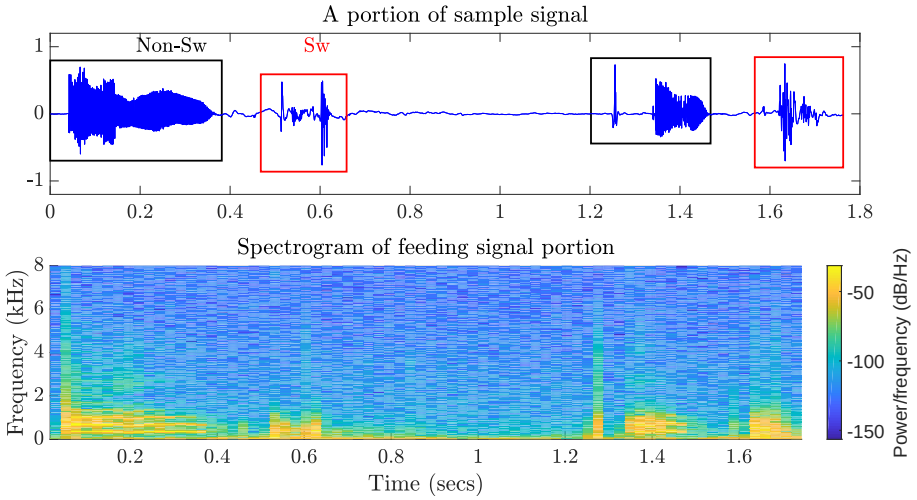
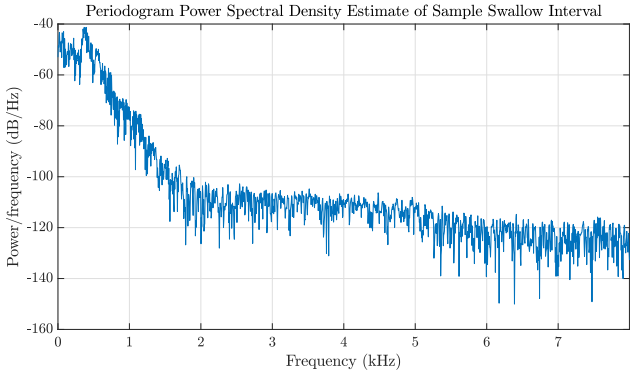
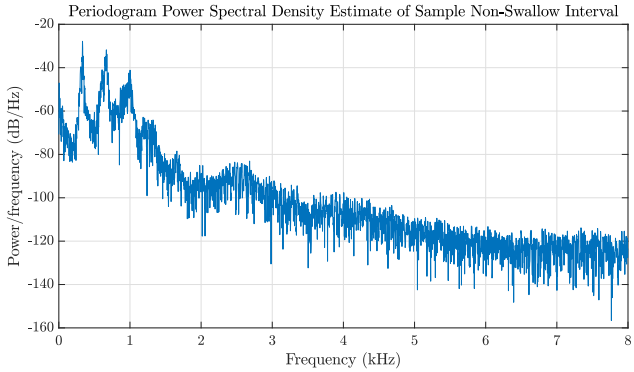


Figure 3.3: Spectrogram of a part of the feeding signal



(a)



(b)

Figure 3.4: (a) Periodogram PSD estimate of an acoustic swallow signal. (b) Periodogram PSD estimate of a non-swallow sound (vowel sound)

CHAPTER 4

METHOD

In this chapter, two different approaches related to the problem of detection of swallowing sound episodes among all other sound activities are discussed.

1. Segmentation followed by Classification
2. Classification followed by Merging

Although they are considered as different approaches, both techniques are based on the pattern recognition methodology and their block diagrams are given in Figure 4.1. However, it is a generic form for different types of classification problems, implying that not all processing stages in the block diagram need to be included. In this study, both approaches were implemented for the same purpose even though pre-processing and post-processing stages differ for each.

Figure 4.1: The flowchart of the pattern recognition methodology

4.1 Segmentation followed by Classification

Considering the graphs and analysis results shown in Chapter 3, a segmentation algorithm is needed for eliminating the silence portions of the acoustic signal. Furthermore, framing and windowing procedures are required to decrease the non-stationary effect of the large interval signal. Thus, features of small frames can be extracted

to be given as an input to the specified classifier. In the end, outputs of classifiers will be used to decide whether a swallow occurs or not with different post-processing algorithms.

In this approach, two different segmentation algorithms will be shown. Later, the feature extraction part will be discussed. Lastly, unsupervised and supervised learning algorithms are explained in detail for the classification part. Generic flowcharts for the training and test phases of this method are depicted in Figure 4.2

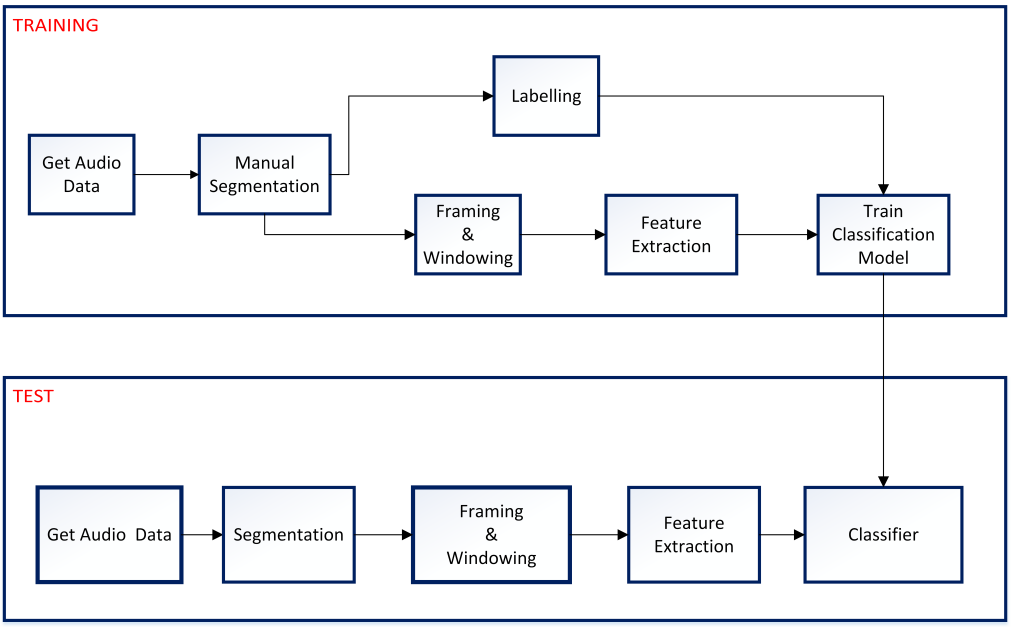


Figure 4.2: Training and the test procedures for the segmentation followed by classification approach

4.1.1 Segmentation

In order to find the swallowing sound activity, the first step may be identifying the intervals of sound activity. For that reason, two segmentation algorithms are designed to detect boundaries of sound activities regardless of swallow or non-swallow events. One of them is based on thresholding the signal energy, whereas the other one utilizes extra features and approaches the problem as a voice activity detection, which is commonly used in the audio signal processing area.

4.1.1.1 Energy Thresholding Based Segmentation

In order to automatically segment boundaries of sound activities, this method utilizes the energy of the consecutive frames with a certain frame length and overlap ratio. Energy of the i^{th} frame, E_i, is computed as

$$E_i = \sum_{n=0}^{N} x_i^2[n] \tag{41}$$

where $x_i[n]$ is the n^{th} sample of i^{th} frame signal of length N. A sample recording and the corresponding energy pattern are given in Figure 4.3.

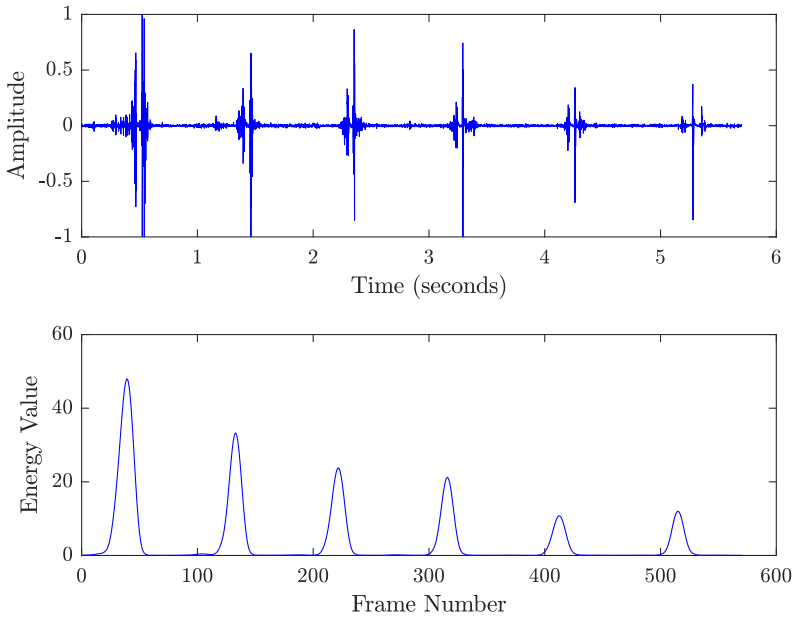


Figure 4.3: Sample recording at the top, corresponding energy plot at the bottom

Sound activity boundaries were detected in two steps; peak detection and boundary detection. In the previous step, the energy signal was passed through third-order one-dimensional median filter to remove noise-like oscillations. After this smoothing process, the peaks of the filtered signal were found. Then, peaks were selected such that the distance between two peaks should not be smaller than a certain threshold. The second elimination method was based on the deletion of the peaks through the threshold extracted from the filtered energy signal. The pipeline for peak detection is given in Figure 4.4. In the following step, the left and the right boundaries of the

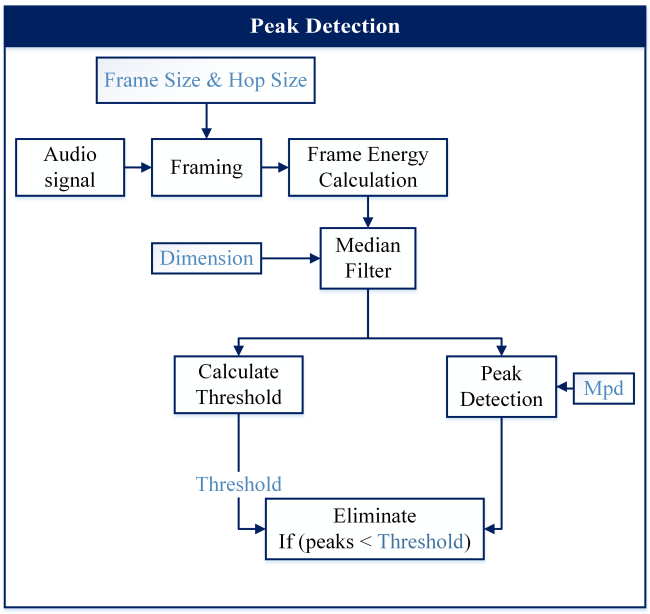


Figure 4.4: Peak Detection Algorithm

segment were determined by navigating to the left and right from the frames of peak points. For this purpose, the energy value of the smaller frames was calculated to extract boundaries more precisely. On both navigating through neighbors and elimination processes, thresholding was utilized. This procedure is given in Figure 4.5.

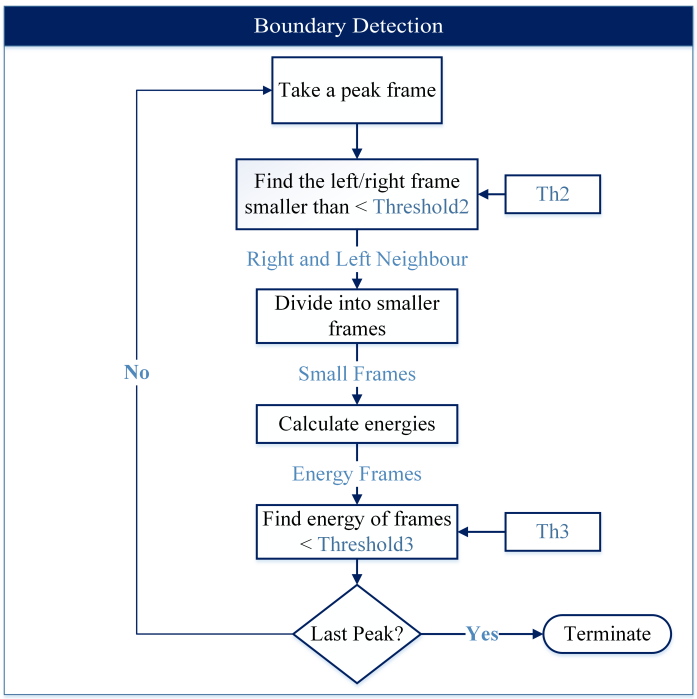


Figure 4.5: Boundary detection algorithm

4.1.1.2 Pattern Recognition Based Segmentation

Voice activity detection (VAD) algorithms are used as pre-processing phases in the speech processing area to determine the presence or absence of human voice. Since the segmentation problem of this study resembles the VAD, suggested algorithms regarding this topic were reviewed [41]. In the end, pattern recognition based solution, which is for classification of the voiced, unvoiced and silence part of the speech was found to be appropriate for this study [42]. However, only two classes, non-silence and silence, were utilized as distinct from the corresponding study since the aim of segmentation was only to eliminate silence portions of the signal.

In this method, firstly, the acoustic feeding signal was divided into smaller frames 10 ms. Secondly, five different audio features were extracted for both training and test parts. Thirdly, in the training part, assuming the distribution of feature observations in the 5-D space as normal, mean vector and covariance matrix of the probability distribution function were computed for both classes. In the test part, the Mahalanobis distances between the feature vectors and the distribution of both classes were calculated. Finally, the feature vector was assigned to the class with minimum distance. The flow

diagram of the segmentation implementation with minor additions and changes to the algorithm used in the paper is shown in Figure 4.6.

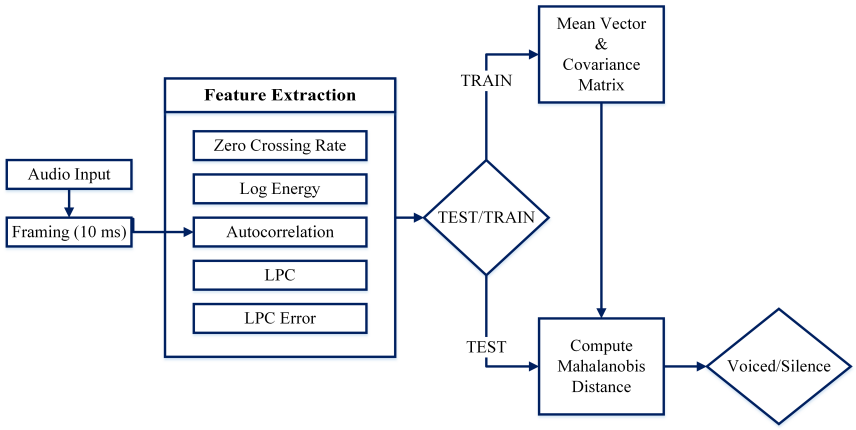


Figure 4.6: Block diagram of the segmentation problem (both train and test parts)

After the framing procedure was applied to the signal, five different measurements were utilized in order to classify corresponding frames as silence or non-silence.

1. **Zero Crossings ($\mathbf{N_z}$)**

   As the feature name suggests, N_z, stands for the number of zero crossings in the frame signal.

2. **Log Energy ($\mathbf{E_{li}}$)**

   The log energy of the i^{th} frame signal, E_{li}, calculated as

   $$E_{li} = 10 \log \left(\frac{1}{N} \sum_{n=1}^{N} x_i^2[n] \right) \tag{42}$$

   where $x_i[n]$ is the n^{th} sample of i^{th} frame signal of length N.

3. **Autocorrelation Coefficient at Unit Sample Delay ($\mathbf{C_1}$)**

   This parameter shows how adjacent samples are correlated with each other. Therefore, if the low-frequency contents dominate the signal, the autocorrelation coefficient for the unit sample delay, C_1, will give a result close to unity implying the higher correlation. C_1 is defined as

   $$C_1 = \frac{\sum_{n=2}^{N} x_i[n] x_i[n-1]}{\sqrt{(\sum_{n=2}^{N} x_i^2[n])(\sum_{n=1}^{N-1} x_i^2[n])}}. \tag{43}$$

4. **The First Coefficient of the Linear Prediction Coding (α_1)**

   Linear Predictive Coding, LPC, is a method generally used in speech processing, speech coding and audio signal processing areas. It is based on the assumption that any sample of the audio signal can be approximated with the linear combination of the past samples [43]. Using the 12-pole LPC, the n^{th} sample of i^{th} frame signal, $x_i[n]$, is approximated as

$$x_i[n] = \sum_{k=1}^{12} \alpha_k x_i[n-k]. \tag{44}$$

   As a result, 12 coefficients were estimated using the covariance method and the first one (α_1) was used as a feature.

5. **Normalized Prediction Error (E_p)**

   This term represents the error of the signal energy in decibels after approximation with the LPC method. The error is calculated as

$$E_p = E_{li} - 10\log\left(\epsilon + \left|\sum_{k=1}^{12} \alpha_k \phi(0,k) + \phi(0,0)\right|\right) \tag{45}$$

   where $\phi(j,k)$ represents the value in the j^{th} row and k^{th} column of the covariance matrix.

In this approach, feature observations were assumed to be normally distributed so training the classifier model involves only calculating the mean vectors and covariance matrices for both classes. In detail, assume that the total number of frames in silence and non-silence intervals are N_0 and N_1, respectively. Therefore, two feature sets, $\boldsymbol{X_0}(N_0 \times 5)$ and $\boldsymbol{X_1}(N_1 \times 5)$, were formed. The mean vector, $\vec{\mu_w}$, and the covariance matrix, $\boldsymbol{\Sigma_w}$ for both classes were computed as

$$\mu_w[k] = \frac{1}{N}\sum_{n=1}^{N_i} X_w(n,k) \qquad \text{(Mean Vector)} \tag{46}$$

$$\boldsymbol{\Sigma_w} = \frac{1}{N}\boldsymbol{X_{cw}^T X_{cw}} \qquad \text{(Covariance Matrix)} \tag{47}$$

$$X_{cw}(j,k) = X_w(j,k) - \mu_w(k) \tag{48}$$

where:

w : class ID ('0' for silence, '1' for voiced)

$\boldsymbol{X_{cw}}$: feature matrix subtracted by the mean feature vector for the i^{th} class

k : feature index.

Testing whether a frame signal belongs to voiced or silence class is based on measuring the Mahalanobis distance between the feature vector of the corresponding frame subtracted from the mean and the probability distribution of the classes. Therefore, the frame signal was assigned to the class, having the distribution with a minimum distance. The Mahalanobis distance between the feature vector, $\vec{x_i} = (x_{i1}, x_{i2}, .., x_{i5})$, and the probability distribution of i^{th} class were calculated as

$$D_i(x_i) = \sqrt{(\vec{x_i} - \vec{\mu_i})^T \Sigma^{-1} (\vec{x_i} - \vec{\mu_i})}. \tag{49}$$

The distance based classifier outputs two different distance values for each frame. However, a post-processing stage is required since the main problem is to separate silence and voiced portions of the related signal. For that reason, the outputs of the classifiers (1's and 0's) were combined with a Finite State Machine (FSM) decision algorithm.

Merging Frames via FSM: In this algorithm, the outputs of the classifier were merged according to certain rules and thresholding methods to detect sound activities. The FSM diagram is shown in Figure 4.7

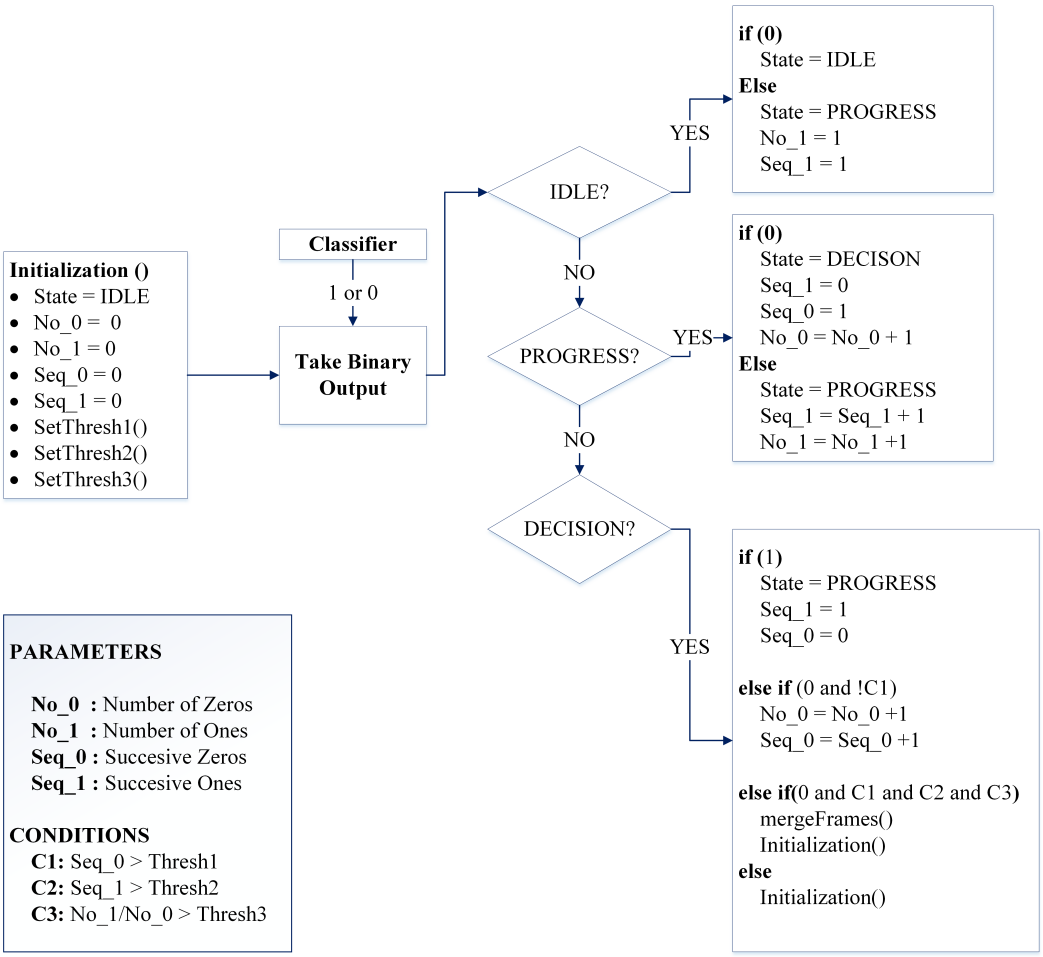


Figure 4.7: Detailed block diagram of the FSM Decision Algorithm

4.1.2 Feature Extraction

A typical audio signal can be used for different purposes by extracting temporal and spectral features. However, features such as zero crossing rate, attributes related to the energy of the signal, etc., are not considered as powerful features for swallow sound classification since the information extracted from the most of the typical time domain features can be extracted in the frequency domain as well. Furthermore, since energy-related features were utilized in the segmentation stage for the elimination of silence and non-silence portions of the signal, they were not considered to be used for distinguishing between swallowing and non-swallowing regions in the vocal parts. Therefore, the frequency domain characteristics of the swallowing sound were analyzed and examined in general.

This section aims to find out the features to be given as inputs for the different types of classifiers. The spectral characteristics of all kinds of sounds were analyzed and examined in detail. Thus, firstly, several frequency domain features of frame signals were extracted and the relation between each frame number and the corresponding feature was observed graphically for both manually labeled swallow and non-swallow sound intervals. Secondly, taking how different combinations of attributes affect the performance of the EM-GMM classifier into account, redundant features were eliminated. In addition to performance, decreasing the time complexity of the feature extraction and classification processes is another concern for this issue.

In this study, swallow sound was assumed to be non-stationary (statistical properties are not constant at any time) like speech. For that reason, the entire feeding signal was divided into small frames so that the assumption that the signal is stationary can be made. In addition, the windowing function was applied to small frames to suppress edge discontinuities and diminish the spectral leakage due to the framing process. A typical hamming window of length N is defined as

$$w[n] = 0.54 - 0.46cos\left(\frac{2\pi n}{N}\right) \qquad n \in \mathbb{N} : [0, N]. \tag{410}$$

A further progress in reducing severe degradation at the edges is the overlapping of short time frames. Thereby, the time resolution of a signal is increased since each part is examined more than once.

Discrete Fourier Transform (DFT) is used to represent the discrete-time signal in the Fourier domain and can be easily computed via digital computers [44]. The DFT can be considered as a common previous step before the extraction of each spectral feature. Including the windowing operation given in (410), L-point DFT of the i^{th} frame signal, F_{x_i}, is calculated a

$$F_{x_i}[k] = \sum_{n=0}^{N-1} x_i[n]w[n]e^{-\frac{2\pi kn}{N}} \qquad k \in \mathbb{N} : [0, L-1]. \tag{411}$$

4.1.2.1 Spectral Centroid

Spectral centroid is the center of mass of the magnitude Fourier transform of an acoustic signal [45]. For an acoustic signal interval having K frames, $K \times 1$ dimensional

feature matrix is constructed. In below, spectral centroid formula of the frame signal, C_{x_i}, is given by

$$C_{x_i} = \frac{\sum_{k=0}^{L-1} k|F_{x_i}[k]|}{\sum_{k=0}^{L-1} |F_{x_i}[k]|}. \tag{412}$$

4.1.2.2 Spectral Spread

Spectral spread (SS) is another feature giving information about the width of the spectrum [46]. In other words, it represents the deviation around the spectral centroid value. Thus, it is necessary to know the value of spectral centroid for the calculation of SS. For an acoustic signal having K frames, $K \times 1$ dimensional feature matrix is constructed. In below, the formula for SS of the frame signal, SS_{x_i}, is given by

$$SS_{x_i} = \frac{\sum_{k=0}^{L-1} (k - C_{x_i})^2 |F_{x_i}[k]|}{\sum_{k=0}^{L-1} |F_{x_i}[k]|}. \tag{413}$$

4.1.2.3 Spectral Flatness

Spectral flatness (SF) is a feature that indicates how tonal or noisy the signal is. Moreover, the ratio of geometric mean to the arithmetic mean of a power spectrum is the traditional definition of it [47]. However, it is possible to see zero magnitudes for some single frequency values in the spectrum, causing the geometric mean equal to zero. In other words, the length of the frame signal is not large enough to have a non-zero value in each frequency bin. Therefore, to calculate the geometric mean of the spectrum, each frequency magnitude of a sub frequency band is summed up rather than using the amplitude of the single frequency value. For an acoustic signal having K frames, $K \times 1$ dimensional feature matrix was constructed. In below, the formula for SF of the frame signal, SF_{x_i}, is given by

$$P_{x_i}[k] = \frac{1}{N f_s} |F_{x_i}[k]|^2 \qquad k \in \mathbb{N} : [0, L-1] \tag{414}$$

$$SF_{x_i} = \frac{\sqrt[L']{\prod_{k=0}^{L'-1} P'_{x_i}[k]}}{\sum_{k=0}^{L'} P'_{x_i}[k]} \qquad k \in \mathbb{N} : [0, L'-1] \tag{415}$$

where $\overrightarrow{P_{x_i}}$ represents the periodogram estimate of the power spectrum. On the other hand, k^{th} element of $\overrightarrow{P'_{x_i}}$ is obtained by summing the amplitudes of $\frac{L}{L'}$ consecutive

frequency bins of $\overrightarrow{P_{x_i}}$.

4.1.2.4 Mel Frequency Cepstral Coefficients

The research and applications of the field of audio and speech processing are mainly developed by considering the human auditory system. The human ear can be considered as a filter concentrated non-uniformly on specific regions of the frequency spectrum. Since the frequency discern skill of the human ear decreases with increasing frequency, the low-frequency region contains more filters than the high one. For the same reason, the perception of the human auditory system cannot linearly evaluate pitch in terms of Hz scale. To express this perception as linear, Mel frequency was described [48]. Mel Frequency Cepstral Coefficients (MFCC) are derived from that logic and dominating the speech and audio processing field for a long time thanks to their ability to represent audio signals in a compact form. The pipeline for calculation of MFCC is depicted in Figure 4.8 Framing and windowing stages are mentioned

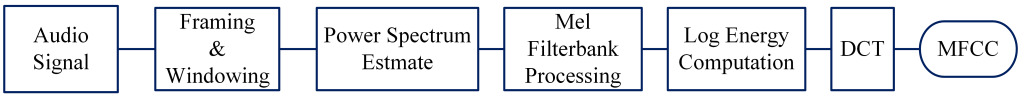


Figure 4.8: Procedure for extracting the MFCC Values

and periodogram estimate of the power spectrum (414) is formulated in the previous sections. Each process other than those mentioned is examined in the following paragraphs [49].

Mel filterbanks can be considered as a cluster of triangular filters which are linearly separated in the Mel frequency scale. Each filter can be represented as a vector of dimension $1 \times (\frac{L}{2} + 1)$, where L is the DFT dimension. In the Figure 4.9, triangular filterbank including 10 filters and $10+2$ boundary points are seen in Hz scale. The corresponding figure is to show the change in the difference between each boundary point of the filterbanks as increasing frequency. It is necessary to declare extreme boundaries for the frequency in terms of Hz to form such a filterbank. After that, lower bound (f_l) and upper bound (f_u) are converted to Mel frequencies, (mel_l, mel_u), with the given formula

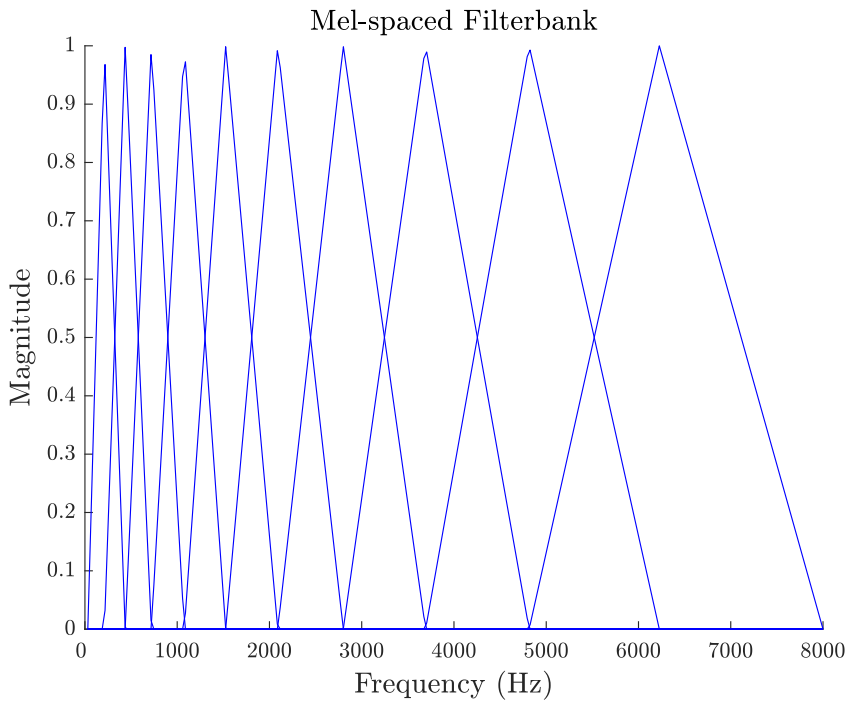


Figure 4.9: Plot of Mel filterbank with 10 filters, when minimum and maximum frequencies are 0 kHz and 8000 kHz respectively.

$$mel_{l,u} = 1127 \log(1 + \frac{f_{l,u}}{700}). \tag{416}$$

After the extreme boundary points are calculated, filter boundary points (linear in Mel scale) are found. Then, each Mel frequency converted back to the frequency scale using the inverse of the equation (416). Finally, m^{th} triangular filter of the filterbank, $\overrightarrow{H_m}$, is created according to the given formula

$$H_m[k] = \begin{cases} 0 & k < 1 \\ \frac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{2(f[m+1]-k)}{(f[m+1]-f[m-1])(f[m+1]-f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases} \quad m \in \mathbb{N} : [0, M-1]. \tag{417}$$

The natural logarithm of energy value at the output of the m^{th} filter, $\overrightarrow{LE_m}$, is then computed with the inner product of the power spectrum estimate and the m^{th} filter

coefficients and is given by

$$LE_m[l] = \log \sum_{l=0}^{\frac{L}{2}+1} P_{x_i}[l] H_m[l]. \tag{418}$$

After the log-energy value vector of size M is obtained, discrete cosine transform (DCT) is applied and MFCC values, $\overrightarrow{CC}$, are calculated as

$$CC[n] = \sum_{k=0}^{M-1} LE_m[k] \cos\left[\frac{\pi}{M}\left(k + \frac{1}{2}\right) n\right] \qquad n \in \mathbb{N} : [0, M-1]. \tag{419}$$

In the end, an M-dimensional MFCC vector was found and the first 13 coefficients were used as features for this study. Therefore, for an acoustic signal having K frames, $K \times 13$ dimensional feature matrix was constructed.

4.1.3 Classification

In this section, different techniques are discussed to classify a segmented interval, as swallow or non-swallow. In the previous sections, the segmentation of non-silence portions and the extraction of several features of each frame of corresponding segmented intervals are explained in detail. Here, the classification of each frame with several learning algorithms and how outputs of multiple frames are combined are told.

4.1.3.1 K-means Clustering

The K-means clustering method is an unsupervised and simple method to solve clustering problem [50]. In other words, if an observation dataset is not labeled, this method will group the unlabeled data with an input parameter K. Let N be the number of observations and K is the number of clusters. The clustering procedure is given in the following steps:

1. Randomly initialize center of clusters, $\boldsymbol{C} = [\vec{c_1}, \vec{c_2}, \ldots, \vec{c_K}]$
2. Calculate euclidean distance between the each sample and the center of clusters.
3. Assign data sample to the cluster whose mean is nearest to it in terms of euclidean distance.

4. Update the mean values of clusters such as

$$\vec{c_i} = \frac{1}{N_i} \sum_{k=1}^{N_i} \overrightarrow{X_i}[k] \tag{420}$$

where i represents cluster ID and $\overrightarrow{X_i}[k]$ is the k^{th} sample of the observation matrix belonging to cluster i and N_i is the total number of samples currently in cluster i.

5. Go to step 2 if convergence is not achieved, otherwise stop the algorithm.

Although K-means clustering is an unsupervised learning algorithm, a decision mechanism was constructed to determine the class of the segmented interval. In order to utilize this method as if supervised learning, clustering operation has been applied separately for each class with the same K value. However, it should be taken into consideration that multiple feature vectors are extracted from each manually labeled interval. Thus, regardless of which interval it comes from, each feature vector was merged according to its class ID. Accordingly, two feature matrices, $\boldsymbol{X_0}(N_0 \times d)$, $\boldsymbol{X_1}(N_1 \times d)$, were formed given that the feature dimension is d. After that, the K-means procedure was applied to both feature matrices to obtain cluster probabilities for each class. The probability of a feature vector given class w (0 or 1) and i^{th} cluster, π_{wi}, is

$$\pi_{wi} = \frac{N_{iw}}{N_w} \tag{421}$$

where N_{iw} is the number of frames belonging to class w , cluster i and N_w is the total number of frames of class w.

Extracting the cluster and class probabilities for both classes can be considered as the training procedure. Let $\boldsymbol{XI} = [\overrightarrow{xi_1}, \overrightarrow{xi_2}, \ldots, \overrightarrow{xi_{NI}}]$, is the feature matrix of the segmented interval with size $NI \times d$. To test whether $\boldsymbol{XI}$ represents a swallow interval or not, the likelihood function of both classes can be calculated as

$$l_w = \sum_{i=1}^{K} \pi_{wi} n_{iw} \tag{422}$$

where n_{iw} is the number of frames assigned to cluster i for the class w. The assignment rule of only one feature vector was to select the cluster giving the minimum Euclidean distance. As a result, $\boldsymbol{XI}$ was appointed as swallow episode if $l_0 > l_1$, or else as non-swallow.

4.1.3.2 Gaussian Mixture Model (GMM)

In the K-means algorithm, features of each frame were assigned to exactly one cluster. However, the possibility of overlapping clusters in the feature space was neglected in this assumption. Also, Euclidean distance to the cluster center is utilized in K-means, yet the clusters may have a non-circular shape. For that reason, clusters are modeled as normally distributed, not just by their mean but also the covariance matrix in the Gaussian mixture model (GMM).

GMMs are the linear combination of the multiple Gaussian components the corresponding probability density function of the model having K mixtures is in the form of [51]

$$p(\overrightarrow{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\overrightarrow{x}|\overrightarrow{\mu_k}, \boldsymbol{\Sigma_k}). \tag{423}$$

Expectation Maximization (EM) is an elegant and robust algorithm that is used for maximum likelihood parameter estimation [52]. GMM is an example of a probabilistic model including parameters such as means, variances and mixing coefficients. Therefore, the aim of Expectation-Maximization for Gaussian Mixture Model (EM-GMM) to maximize the likelihood function in terms of GMM parameters.

In this study, both classes were assumed to be having K Gaussian mixtures. For class w, the algorithm steps for maximizing the log-likelihood function is given below.

1. Initialize means, $\overrightarrow{\mu_k^w}$, covariance matrices, $\boldsymbol{\Sigma_k^w}$, and mixing coefficients, π_k^w
2. Find the responsibility of each sample ($\overrightarrow{x^w}[n]$) in class w according to formula

$$\gamma^w[n,k] = \frac{\pi_k^w \mathcal{N}\left(\overrightarrow{x^w}[n]|\overrightarrow{\mu_k^w}, \boldsymbol{\Sigma_k^w}\right)}{\sum_{m=1}^{K} \pi_m^w \mathcal{N}(\overrightarrow{x^w}[n]|\overrightarrow{\mu_m^w}, \boldsymbol{\Sigma_m^w})} \tag{424}$$

3. Update the means, variances and mixing coefficients with the help of calculated responsibilities

$$\overrightarrow{\mu^w}_{k,new} = \frac{1}{N_k^w} \sum_{n=1}^{N} \gamma^w[n,k] \overrightarrow{x^w}[n] \tag{425}$$

$$\boldsymbol{\Sigma_{k,new}^w} = \frac{1}{N_k^w} \sum_{n=1}^{N} \gamma^w[n,k] \left(\overrightarrow{x^w}[n] - \overrightarrow{\mu^w}_{k,new}\right) \left(\overrightarrow{x^w}[n] - \overrightarrow{\mu^w}_{k,new}\right)^T \tag{426}$$

$$\pi_{k,new}^{w} = \frac{N_k^w}{N^w} \tag{427}$$

where N^w is the total number of feature vectors in class w and N_k^w is given as

$$N_k^w = \sum_{n=1}^{N^w} \gamma^w[n,k] \tag{428}$$

4. Check whether log-likelihood function given below converged or not.

$$log(p^w(\overrightarrow{x^w})) = \sum_{n=1}^{N_w} log\left[\sum_{k=1}^{K} \pi_k^w \mathcal{N}\left(\overrightarrow{x^w}[n]|\overrightarrow{\mu_k^w}, \boldsymbol{\Sigma_k^w}\right)\right] \tag{429}$$

5. Go to step 2 if convergence is not achieved, otherwise stop the algorithm.

Decision Rules for EM-GMM: Maximizing the likelihood function by estimating the unknown parameters of the Gaussian mixtures can be named as the training phase of this algorithm. For both classes, K means, covariance matrices and mixing coefficients were estimated. After that, two decision rules were proposed to merge the outputs of the classifier. Here, output refers to the likelihood of only one audio frame signal.

1. **Sum of log-likelihoods:** The feature vectors inside a segmented interval were assumed to be independent. Based on this assumption, the product of likelihood values of each frame, or sum of the log-likelihood values were computed for both swallow and non-swallow cases. Finally, the input signal having N_f frames was classified by comparing the sum of logarithms.

$$log\left(p_{total}^{w}(\boldsymbol{X}|\boldsymbol{\Sigma^w}, \overrightarrow{\mu^w}, \boldsymbol{\pi^w})\right) = \sum_{n=1}^{N_f} log\left(\sum_{k=1}^{K} \pi_k^w \mathcal{N}\left(\overrightarrow{x^w}[n]|\overrightarrow{\mu_k^w}, \boldsymbol{\Sigma_k^w}\right)\right). \tag{430}$$

   If the total sum of log-likelihood of swallow (non-swallow) class was higher, the input signal was classified as swallow (non-swallow).

2. **Majority Voting:** Frame by frame comparison between the two likelihood values of a single frame of an episode and majority voting may result in a more accurate decision. Therefore, for a segmented interval containing N_f frames,

N_f comparisons were done. If the total number of frames favoring swallow (non-swallow) class was higher than the non-swallow (swallow) class, the input signal was classified as swallow (non-swallow).

Algorithm 1, 2, 3 show the pseudo codes for the units such as building classification models and testing the extracted models using the GMMs.

Algorithm 1: Creation of Feature Observation Matrix in GMM

input : Collection of audio records K
output: $featMat$, a matrix that contains audio features of frames
for $\forall i \in K$ **do**
    $K_i \leftarrow$ `extractLabelInfo(`i`)` `// Extract intervals and labels`
    for $\forall frame \in K_i$ **do**
        Compute MFCC values of $frame$
        Compute spectral Centroid of $frame$
        Concatenate features
        Append feature vector to $featMat$
    end
end

Algorithm 2: Building a Classification Model with EM-GMM

input : Unnormalized observation matrix $featMat$
output : Classifier model, $gmmModel$
parameter: Tolerance value related to convergence rule, $tolValue$
$(coeffs, scores) \leftarrow$ `PCA(`$featMat$`)`
for *i=1,2 (2 classes)* **do**
    Find means of each mixture with the help of K-means algorithm
    Find corresponding labels for each mixture
    for *k=1,2 ... K (# of gaussian mixtures)* **do**
        Assign the mean values as found in K-means
        Extract each feature vector having the same mixture label *k*
        Concatenate corresponding feature vectors and form a matrix
        Initialize the covariances and mixing coefficients
    end
end
for *i = 1,2 (for both swallowing and non swallowing frames)* **do**
    Convergence Rule : $ll_t - ll_{t-1} < |(ll_t)|tolValue$
    while *convergenceIsNotSatisfied()* **do**
        E step :Calculate responsibilities
        M step :Re-estimate the parameters using the current responsibilities
        Evaluate the ll
        Check for convergence of ll
    end
    $gmmModel_i \leftarrow$ estimated parameters
end
Function `PCA(`$obsMatrix$`)`
    $meanObsMatrix \leftarrow$ Mean of observation matrix
    $zeroMeanObsMatrix = obsMatrix - meanObsMatrix$
    $coeffs \leftarrow$ Principal component coefficients of $zeroMeanObsMatrix$
    $scores \leftarrow$ Projection of $zeroMeanObsMatrix$ onto the principal components
    return $meanObsMatrix, coeffs, scores$

Algorithm 3: Testing the EM-GMM Classifier

input : Collection of audio records, K, Trained GMM Models, $gmmModels$
output: $decisionVec$, holding the interval decisions

for $\forall i \in K$ **do**
    $K_i \leftarrow$ `segmentation(i) // (`
    Group of segmented interval) **for** $\forall j \in K_i$ **do**
        for $\forall k \in j$ **do**
            Compute MFCC values of k^{th} frame of j^{th} interval
            Compute spectral centroid of k^{th} frame of j^{th} interval
            $featVec \leftarrow$ Concatenate features
            $ll_{sw}[k], ll_{nsw}[k] \leftarrow$ `computeLikelihood(`$gmmModels, featVec$`)`
        end
        `// Majority Voting,` K_ij `represents one interval`
        if *Majority* == sw **then**
            $K_{ij} \leftarrow$ swallow interval
            //
        else
            $K_{ij} \leftarrow$ non-swallow interval
        end
        `// Sum of log-likelihoods`
        if ($\sum ll_{sw} > \sum ll_{nsw}$) **then**
            $K_{ij} \leftarrow$ swallow interval
        else
            $K_{ij} \leftarrow$ non swallow interval
        end
    end
end

4.1.3.3 Gaussian-HMM

In the previous Gaussian mixtures based classification method, sequential information was not considered. Each small frame of a segmented interval was evaluated independently and the decision was given without thinking any temporal dependency between consecutive frames. However, swallowing is a sequential function satisfying the proper coordination of muscles of mouth, palate, pharynx, larynx, esophagus [8]. Hence, it can be modeled as a Markov Process containing N different states. Since the states cannot be observed physically, the system is described as Hidden Markov Model [53] for this problem. An example of the lattice representation of HMM with 3 states is illustrated in Figure 4.10. $\vec{t} = [1, 2, \ldots, T]$ stands for the time sequence vector, while $\vec{O} = [O_1, O_2, \ldots, O_T]$ are the observations. Furthermore, S_1, S_2 and S_3 are the three unobserved states of the HMM. The transition probability from the state i to state j is represented as a_{ij}.

In a discrete HMM, the probability of i^{th} observation within the state j at any time t is depicted as b_{ij}. However, when it is not possible to describe observations discretely, HMM is used with continuous observation densities. Since there are no solid

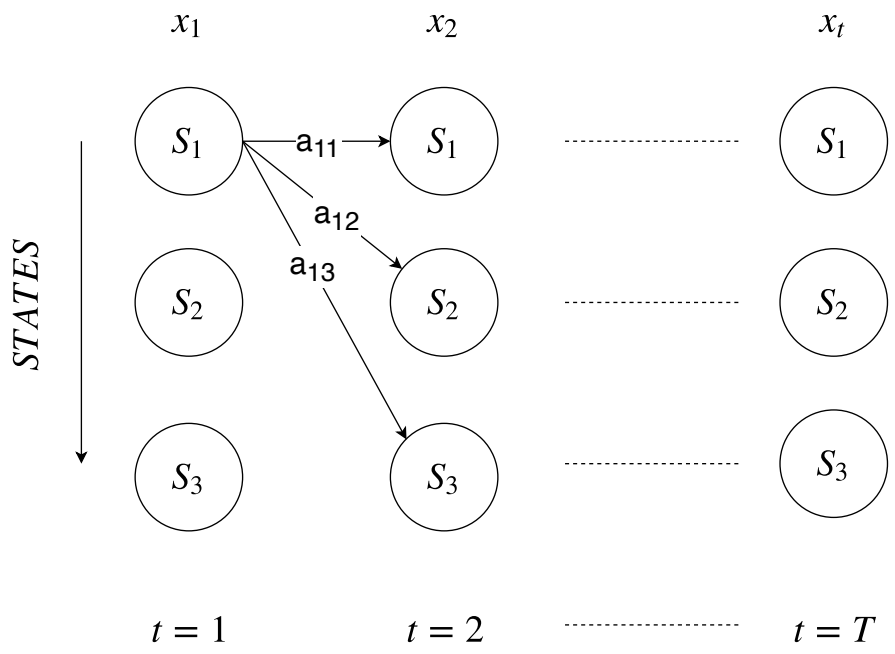


Figure 4.10: Lattice representation of an HMM with 3 states and T observations

observation definitions for the swallow function, observation densities are computed by assuming the distribution of each state as Gaussian. Therefore, given state j, the density of any feature vector, $\boldsymbol{x}$, is calculated as

$$b_j(\vec{x}) = p(\vec{x}|S_j) = \mathcal{N}(\vec{x}; \vec{\mu_j}, \boldsymbol{\Sigma_j}). \tag{431}$$

In an HMM, three main problems are discussed, namely, evaluation, recognition and training. The evaluation problem is merely finding the probability of an observation sequence $O = [O_1, O_2, \ldots, O_T]$ in which initial state probabilities, transition probability matrix and observation densities are known. In the recognition problem, the goal is to find the best state sequence, $\vec{Q} = [q_1, q_2, \ldots, q_T]$, of a given model and the likelihoods of frames at any time t. The third problem is the estimation of the HMM parameters, which are transition probability matrix, $\boldsymbol{A}$, initial state probabilities of states, $\pi = [\pi_1, \pi_2, \ldots, \pi_N]$ (N is the number of states) and unknown parameters of Gaussian distribution.

The most significant difference between Gaussian-HMM and the GMM for this study can be the presence of the transition probability matrix. Therefore, as in the case of GMM, two distinct Gaussian-HMMs were constructed to interpret the problem as

binary classification.

In the evaluation or the test part, the likelihood of observations, $\boldsymbol{X} = [\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_T}]$, given the model, λ, i.e, $P(\boldsymbol{X}|\lambda)$ is calculated. If the state sequence, $\vec{Q} = [q_1, q_2, \ldots, q_T]$, is also given, then the likelihood of continuous observation sequence can be described as

$$p(\boldsymbol{X}|\vec{O}, \lambda) = \prod_{t=1}^{T} p(\overrightarrow{x_t}|q_t). \tag{432}$$

On the other hand, the probability of the state sequence, $\vec{Q}$, will be in the form

$$P(\vec{Q}|\lambda) = \pi_{q_1} \prod_{t=2}^{T} \mathrm{a}_{q_{t-1}q_t} \tag{433}$$

Thus, the probability of Q and X at the same time for the given model will be

$$p(\boldsymbol{X}, \vec{Q}|\lambda) = p(\boldsymbol{X}|\vec{Q}, \lambda)P(\vec{Q}|\lambda) = \pi_{q_1} \prod_{t=1}^{T} p(\overrightarrow{x_t}|q_t) \prod_{t=2}^{T} \mathrm{a}_{q_{t-1}q_t}. \tag{434}$$

The above term is valid for a specific state sequence. Yet, the likelihood of continuous observations should be calculated for each state sequence combination and then summed which can be shown as

$$P(\boldsymbol{X}|\lambda) = \sum_{Q} p(\boldsymbol{X}, \vec{Q}|\lambda) = \sum_{Q} \pi_{q_1} \prod_{t=1}^{T} p(\overrightarrow{x_t}|q_t) \prod_{t=2}^{T} \mathrm{a}_{q_{t-1}q_t}. \tag{435}$$

As it can be inferred from (435), the number of state sequence combination for N states and T timestamps equals to N_T and for each state sequence, $2T$ calculations are required, thus yielding $2TN^T$ arithmetic operations in total. This method is, unfortunately, inapplicable due to high complexity. To lower complexity, the forward procedure is utilized.

Forward Procedure: In this methodology, the forward variable, $a_t(j)$, is defined as the likelihood that the partial observation sequence, $\boldsymbol{X} = [\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_T}]$, occurs and the state at time t equals to S_j. After setting the initial values, $a_t(j)$ is solved as follows:

$$a_t(j) = \left(\sum_{i=1}^{N} a_{t-1}(i)\mathrm{a}_{ij} \right) p(\overrightarrow{x_t}|q_t = S_j) \tag{436}$$

The above equation means that the likelihood of the partial observation sequence at time t for a given state j can be derived from each forward variable of previous states

and the given transition probabilities between states. In the Figure 4.11, updating procedure is illustrated. At the end ($t = T$), the likelihood of the entire observation sequence can be found as

$$P(\boldsymbol{X}|\lambda) = \sum_{i=1}^{N} a_T(i). \tag{437}$$

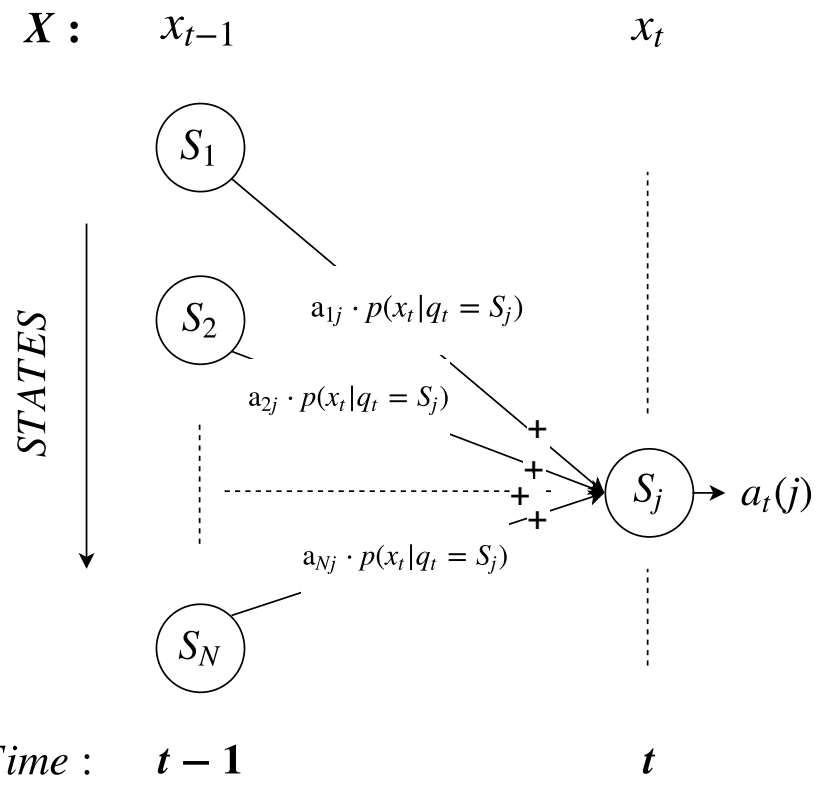


Figure 4.11: Computation of the forward variable from time $t-1$ to t

Thanks to forward procedure, the number of arithmetic operations (additions, multiplications) decreases to N^2T. Remember that the brute-force approach (435) requires on the order of $2TN^T$ calculations. More clearly, for $N = 3$ and $T = 10$, the number of operations in the forward method is 90, whereas on the order of 10^6 operations are needed in the former method.

In this method, previous parts such as segmentation, dividing into smaller frames and feature extraction are the same with those in the EM-GMM algorithm. Similarly, two Gaussian-HMMs were created for each class. Here, the number of frames was considered as the number of timestamps. Assuming that the Gaussian parameters for

each state were known and the transition probability matrix was given, the likelihood of continuous observation densities for both classes were computed applying the forward procedure algorithm. Finally, segmented sound activity was assigned to class having a larger likelihood value at the output.

Backward Procedure: This process is required to calculate the probability of being in state S_j at time t given the entire observation densities. Remember that in equation 436, $a_t(j)$ represents the likelihood of the same model, state and the time-stamp but given partial observation densities. Thus, the backward variable, $\beta_t(j)$, likelihood of being in state S_i at time t given the $\boldsymbol{X} = [\overrightarrow{x_{t+1}}, \overrightarrow{x_{t+2}}, \ldots, \overrightarrow{x_T}]$ is defined as

$$\beta_t(i) = \left(\sum_{j=1}^{N} \beta_{t+1}(j) \mathrm{a}_{ij} \right) p(\overrightarrow{x_{t+1}} | q_{t+1} = S_i). \tag{438}$$

After the forward and backward variables are found for the given state and time, the probability of being in state S_i at time t given the entire observation densities is defined as

$$\begin{aligned} \gamma_t(i) &= \frac{a_t(i)\beta_t(i)}{p(\boldsymbol{X}|\lambda)} \\ &= \frac{a_t(i)\beta_t(i)}{\sum_{i=1}^{N} a_t(i)\beta_t(i)}. \end{aligned} \tag{439}$$

In addition, given the model and continuous observation densities, the probability of being in state S_i at time t and in state S_j at time $t+1$ can be calculated as

$$\varepsilon_t(i,j) = \frac{a_t(i)\beta_{t+1}(j)\mathrm{a}_{ij} p(\overrightarrow{x_{t+1}} | q_{t+1} = S_j)}{p(\boldsymbol{X}|\lambda)}. \tag{440}$$

The Baum-Welch re-estimation aims to find the model assigning the training data the maximum likelihood [54]. As no analytic solution exists, a special case of EM algorithm is utilized. The above two terms will be updated in the expectation part while the mean and covariance matrices belonging to states will re-estimated in the maximization part. In this study, the number of states was assumed to be equal for both classes. Therefore, the same procedure was applied for each class. The algorithm is expressed in the following steps.

1. Initialize the initial state probabilities, π_i, means, $\boldsymbol{\mu_i}$, covariance matrices, $\boldsymbol{\Sigma_i}$, transition probability matrix A.

2. **Expectation:** Calculate $\gamma_t(i)$ and $\varepsilon_t(i,j)$ according to (439) and (440).

3. **Maximization:** Update the initial state probabilities means, covariance matrices and transition probability matrix.

$$\pi_i = \frac{\gamma_1(i)}{\sum_{j=1}^{N}\gamma_1(j)} \qquad \mathrm{a}_{ij} = \frac{\sum_{t=1}^{T}\varepsilon_t(i,j)}{\sum_{t=1}^{T}\gamma_t(i)}$$

$$\vec{\mu_i} = \frac{\sum_{t=1}^{T}\gamma_t(i)\vec{x_t}}{\sum_{t=1}^{T}\gamma_t(i)} \qquad \boldsymbol{\Sigma_i} = \frac{\sum_{t=1}^{T}\gamma_t(i)(\vec{x_t}-\vec{\mu_i})(\vec{x_t}-\vec{\mu_i})^T}{\sum_{t=1}^{T}\gamma_t(i)}$$

4. Check whether convergence is achieved, otherwise go to step 2.

One of the most important issues that should be taken into consideration in a re-estimation based algorithm is initialization. Because the algorithm may be stuck in the local maximum value of the likelihood depending on the initial parameters. To avoid this, firstly, the centroid values of feature vectors for each state were obtained via K-means clustering. After that, independent from the HMM, EM-GMM was applied to estimate Gaussian mixture parameters, namely, means and covariance matrices. Then, each training sample was assigned to a state according to their likelihoods found by the EM-GMM algorithm. On the other hand, the probability of transition from state S_i to S_j,i.e, $\mathrm{a}_{ij,init}$ was initialized as

$$\mathrm{a}_{ij,init} = \frac{\text{Total number of transitions from state } S_i \text{ to } S_j}{\text{Total number of being in state } S_i}. \tag{441}$$

In summary, both manually labeled swallow and non-swallow activities divided into small frames were given to two distinct HMMs as training data. After that, since the space of the observation signals related to each state was modeled as Gaussian probability density function, the unknown parameters of it were estimated with the help of the Baum-Welch Algorithm for each class. With the help of this re-estimation procedure, state transition probability matrix and initial state probabilities were determined as well. On the other hand, in the test part, segmented intervals were given as input to trained HMMs to calculate the likelihood of the entire sequence. Finally, segmented sound activity was assigned to class having a larger likelihood value at the output.

In Algorithm 4, 5, 6, modules, namely, creation of feature matrices, training and testing the Gaussian-HMM model procedures are depicted.

Algorithm 4: Creation of Cell Array of Feature Matrices for one class

input : Collection of audio records K

output: $featMatCellt$, cell array of matrices containing multiple frame features of intervals

Initialize $featMatCell$ as and empty cell
for $\forall i \in K$ **do**
    $K_i \leftarrow$ `extractLabelInfo(`i`)` `// Extract intervals and labels`
    Initialize $intervalFeats$ of K_i as an empty matrix. **for** $\forall frame \in K_i$ **do**
        Compute MFCC values of $frame$
        Compute spectral Centroid of $frame$
        Concatenate features and append to $intervalFeats$
    end
    Append $intervalFeats$ to $featMatCell$
end
return $featMatCell$

Algorithm 5: Training the Gaussian HMM model for one class

input : $featMatCellt$, cell array of matrices containing multiple frame features of intervals, K, number of states, $iterNo$

output: $ghmmModel$, state transition matrix, initial probabilities and gaussian parameters

`// Estimate gaussian parameters of corresponding class for each state`
$params \leftarrow$ `emgmm(`$featMatCell, K$`)`
$stateTransitionNo \leftarrow$ zero matrix of dimension $K \times K$
for $\forall featMat \in featMatCell$ **do**
    Initialize $states$ as an empty vector
    for $\forall frame \in featMat$ **do**
        Initial assignment of $frame$ to a state according to $params$
        Append assigned state to $states$
    end
    `// For Initial Estimate of State Transition Matrix`
    $stateTransitionNo \leftarrow$ `updateStateTransitions(`*states*`)`
end
$A_0 \leftarrow$ `estimateStateTransitionMatrix(`*stateTransitionNo*`)` `// Initial Estimate`
for $i \in [1, iterNo]$ **do**
    `// Expectation Step`
    for $\forall featMat \in featMatCell$ **do**
        Find likelihoods of each feature inside $featMat$
        Forward and Backward Procedure
    end
    `// Maximization Step`
    Update gaussian parameters `//` Σ, μ `and` π `of each state`
    Update state transition matrix
end
return $featMatCell$

Algorithm 6: Testing the Gaussian-HMM Classifier

input : Collection of audio records, K, Trained GMM Models, $gmmModels$, State Transition Matrices, A, Initial State Probabilities, π

output:

for $\forall i \in K$ **do**
    $K_i \leftarrow$ `segmentation(`i`)` `// (`
    Segmented intervals) **for** $\forall j \in K_i$ **do**
        for $\forall k \in j$ **do**
            Compute MFCC values of k^{th} frame of j^{th} interval
            Compute spectral centroid of k^{th} frame of j^{th} interval
            $featVec \leftarrow$ Concatenate features
            Append $featVec$ to feature matrix of K_ij interval
            $pdfOutputsSw[k] \leftarrow$ `computeLikelihood(`$ghmmModel_{Sw}, featVec$`)`
            $pdfOutputsSw[k] \leftarrow$ `computeLikelihood(`$ghmmModel_{NSw}, featVec$`)`
        end
        $likelihoodSw \leftarrow$ `viterbiDecode(`$pdfOutputsSw, \phi_{init}, A_{sw}$`)`
        $likelihoodNSw \leftarrow$ `viterbiDecode(`$pdfOutputsNSw, \phi_{init}, A_{NSw}$`)`
        `// Sum of log-likelihoods`
        if ($likelihoodSw > likelihoodNSw$) **then**
            $K_{ij} \leftarrow$ swallow interval
        else
            $K_{ij} \leftarrow$ non swallow interval
        end
    end
end

4.1.3.4 Support Vector Machines

Whenever applicable, two-class problems can be solved by finding a decision boundary (a hypersurface) in the feature space. In SVM, elements of the training data that are nearest to the separating hyperplane are called support vectors and the main goal is to maximize the distance between support vectors and the decision boundary. Support vectors can be considered as the training samples, which are the most challenging samples for classification [55]. That is why they define the location of the decision boundary. How SVM approaches a linearly separable two class problem is illustrated in Figure 4.12.

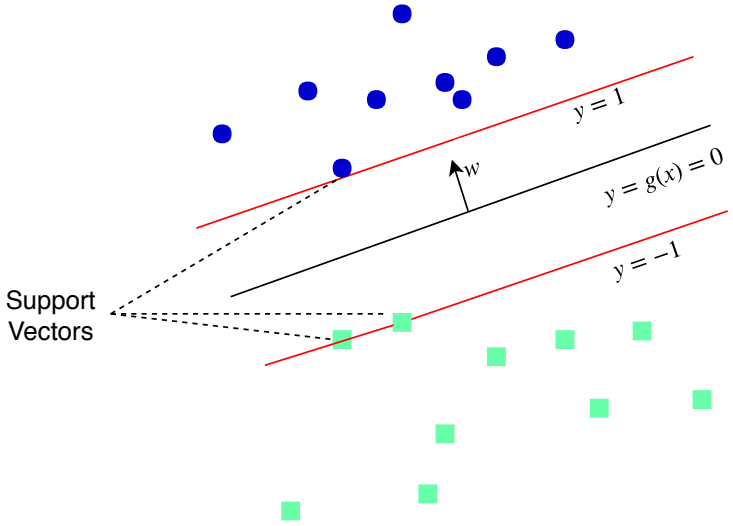


Figure 4.12: The 2-D scheme of the SVM classification. Black line is the decision boundary (Optimal Separating Hyperplane), red lines are called positive and negative hyperplanes, w is the normal vector to the decision boundary, blue circles and green squares stand for two classes.

The $g(\vec{x})$ can be considered as a discriminant function and used to form a decision mechanism such that the positive class should be assigned to sample $\vec{x_i}$, if $g(\vec{x_i}) > 0$ is satisfied or vice versa. The form of $g(\vec{x})$ is given by

$$g(\vec{x}) = \vec{w}^T \vec{x} + w_0 \tag{442}$$

where w is the weight vector and w_0 represents the bias term. In Figure 4.13, the distance between a sample, $\vec{x}_i$, and the decision boundary line (2-D feature space),

$y = 0$, is shown and calculated as

$$\begin{aligned} d &= \frac{g(\vec{x})}{||\vec{w}||} \\ &= \frac{\vec{w}^T\vec{x} + w_0}{||\vec{w}||} \end{aligned} \tag{443}$$

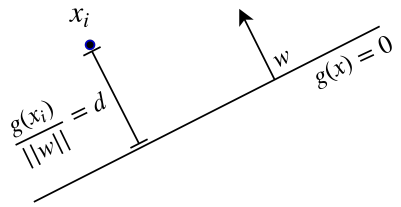


Figure 4.13: View of the distance between 2-D sample $\vec{x_i}$ and decision boundary

The discriminant function, $g(\vec{x})$, can be scaled by a parameter k such that distance between the nearest point, $\vec{x_i}$, and the decision boundary will be $\frac{1}{||\vec{w}||^2}$.

$$\vec{w}^T\vec{x} + w_0 = 0 \implies k\vec{w}^T\vec{x} + kw_0 = \vec{w_{up}}^T\vec{x} + w_{0,up} = 0 \tag{444}$$

$$\frac{\vec{w_{up}}^T\vec{x_i} + w_{0,up}}{||\vec{w_{up}}||^2} = \frac{1}{||\vec{w_{up}}||^2} \tag{445}$$

Optimal decision boundary will be obtained as a result of optimization by maximizing the distance of closest samples which can be shown as

$$\begin{aligned} &\underset{w_{up}}{\text{maximize}} \quad \left(\frac{1}{||\vec{w_{up}}||^2}\right) \text{ or } \underset{w_{up}}{\text{minimize}} \left(||\vec{w_{up}}||^2\right) \\ &\text{subject to} \quad y_i\left(\vec{w_{up}}^T\vec{x_i} + w_{0,up}\right) \geq 1,\ i = 1, \ldots, N. \end{aligned} \tag{446}$$

where $y_i = \{1, -1\}$ is the assigned label of the sample x_i and N is the total number of training sample. It can be inferred from (446) that, the term, $\left(\vec{w_{up}}^T\vec{x_i} + w_{0,up}\right)$, will be negative when the label $y_i = -1$ and positive when $y_i = +1$ if the problem is linearly separable. However, it is not possible to draw a hyperplane to separate two classes in most of the classification problems as the real-world data is generally not linearly separable. For that reason, a generalized discriminant function, $g'(\vec{x})$, is defined as

$$g'(\vec{x}) = \vec{w}^T\phi(\vec{x}) + w_0 \tag{447}$$

where $\phi(\vec{x})$ is the feature map function helping non-linearly separable data to become separable with the help of a linear hyperplane in a higher dimension. Also, although a standard SVM looks to find a margin split all the positive and negative training data, this may lead to a modeling error such as overfitting. In this case, the misclassification rate for train data can be zero, while the error rate of the test data can be found to be quite high. Therefore, to obtain better classification performance on test data, the margin SVM seeks to can be increased. A trade-off parameter, C, is a penalty factor for misclassifying the training instances, the lower C, the higher the final training error. On the other hand, the generalization power of the classifier will increase [56]. Given the trade-off parameter, C, the primal problem of optimization will be

$$\begin{aligned} &\underset{\vec{w_{up}}}{\text{minimize}} \quad \left(||\vec{w_{up}}||^2 + C \sum_{i=1}^{N} \varepsilon_i \right) \\ &\text{subject to} \quad y_i \left(\vec{w_{up}}^T \phi(\vec{x_i}) + w_{0,up} \right) \geq 1 - \varepsilon_i, \; i = 1, \ldots, N. \end{aligned} \tag{448}$$

In order to find a solution for the optimization problem that is constrained to at least one equality or inequalities, the Lagrange Multiplier method is used. Then, applying the Karush-Kuhn-Tucker conditions [56, 57], the dual optimization problem will be

$$\begin{aligned} &\underset{a_i \geq 0}{\text{maximize}} \quad \left(\sum_{i=0}^{N} a_i - \frac{1}{2} \sum_j \sum_k a_j a_k y_j y_k \phi(\vec{x})^T \phi(\vec{x}) \right) \\ &\text{subject to} \quad 0 \leq a_i \leq C \text{ and } \sum_i a_i y_i = 0, \; i = 1, \ldots, N. \end{aligned} \tag{449}$$

The above equation implies that the mapping function, $\phi(\vec{x})$, only occurs in pairs $(\phi(\vec{x_i}), \phi(\vec{x_j}))$ in optimization problem. Therefore, a coefficients will be learned after the dot products of pairs of the training data are calculated. Most of the a_i values will be zero, so the dual form has no disadvantage over the primal problem due to the high number of arithmetic operations. The ones having non-zero values will be named as support vectors. The inner product of mapping functions is defined as the kernel function [58], K, which can be shown as

$$K(\vec{x_i}, \vec{x_j}) = \phi(\vec{x_i})^T \phi(\vec{x_j}) \tag{450}$$

The dimension of the mapped space does not matter since the inner products are utilized. Thus, the kernel function should be selected such that it can be written as

the form of the inner product of pair mapped vectors (450). Also, the classifier can be re-written in the dual form in terms of kernel function as

$$g'(\vec{x}) = \sum_{i=1}^{N} a_i y_i K(\vec{x_i}, \vec{x_j}) + b \tag{451}$$

Some popular kernels are listed below:

$$K(\vec{x_i}, \vec{x_j}) = (\vec{x_i}^T \vec{x_j}) \qquad \text{Linear Kernel} \tag{452}$$

$$K(\vec{x_i}, \vec{x_j}) = (\vec{x_i}^T \vec{x_j})^2 \qquad \text{2-D polynomial Kernel} \tag{453}$$

$$K(\vec{x_i}, \vec{x_j}) = e^{-\gamma ||\vec{x_i} - \vec{x_j}||^2} \qquad \text{Radial Basis Function (RBF) Kernel} \tag{454}$$

Similar to previous approaches, in this study, training and classification pipelines were constructed at the beginning. In the training part, labeled swallow/non-swallow intervals were divided into smaller frames as in the EM-GMM and Gaussian-HMM cases. Then, the normalization procedure was applied to the extracted features of both classes. In the end, the model was trained with normalized features. In other words, support vectors and corresponding a_i values were found (449). In the classification or test part, extracted features were firstly normalized and then given to classifier (451). In general, SVM is a binary classifier that outputs the decision label. However, the majority of the classification problems requires the posterior probabilities as well. For that reason, Platt et al. proposed an algorithm mapping binary decisions into probability scores with an additional sigmoid function [59].

After obtaining the posterior probabilities at the output, a decision mechanism, which is similar to the sum of the log-likelihood estimates approach used in (430) was constructed. To assign the segmented sound interval as swallow or non-swallow, the logarithms of swallow and non-swallow probabilities were summed assuming the independence between frames. If the difference between sums of two classes were higher than a certain threshold, the class was selected to be as swallow which can be shown

$$\sum_{i=1}^{N_s} \log\left(P(\vec{x_i}|Sw)\right) > \sum_{i=1}^{N_s} \log\left(P(\vec{x_i}|NonSw)\right) + THR \implies \text{Swallow Interval.} \tag{455}$$

4.2 Classification followed by Merging

In the first approach (Segmentation followed by Classification), the audio data sequence was firstly segmented with distinct segmentation algorithms as a pre-processing stage. Secondly, extracted intervals were divided into frames and each frame was classified as swallow or non-swallow. Finally, different decision mechanisms were constructed to make a classification in terms of the segmented interval. However, in this case, no pre-processing algorithm was applied to the signal. Instead, the entire data sequence was split into frames and the features of each were extracted. Then, extracted features were given as inputs to the classifier. In order to detect swallow boundaries, the outputs of the classifier were combined with different merging algorithms. Therefore, it can be considered as a method in which the segmentation and decision mechanisms are made after the classifiacation process. For both training and test phases, the generic flowchart of the proposed approach is given in Figure 4.14

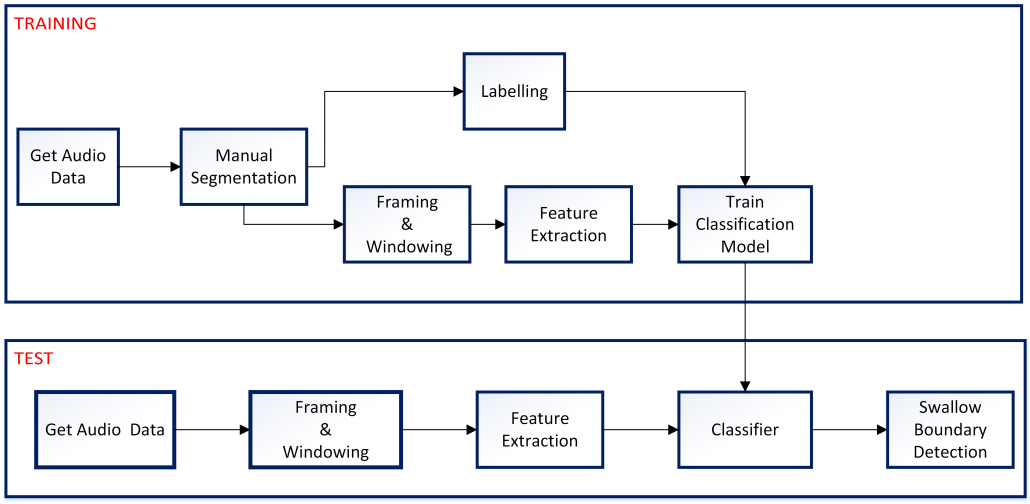


Figure 4.14: Training and the test procedures for the Classification followed by Merging Approach

The framing and windowing procedures are the same as the previous approach, yet two different SVM-based classifiers were used with similar merging algorithms. Thus, only the implementation processes of the classifier algorithms are discussed in this section. In both ways, the trained classification model returns scores, the posterior probabilities, of specified classes for each frame. Then, output scores were examined

to extract swallow episodes.

4.2.1 Binary SVM Classifier

In this case, the swallow represents one class, while the combination of non-swallow and silent parts constitutes the other class. In addition to manually labeled swallow and non-swallow sections, the features of silence parts were also extracted. Remember that, if an audible episode cannot be labeled as swallow or non-swallow, it is marked as ambiguous. Thus, no extra effort is needed for labeling them as all the unlabeled part of the signal represents the silence portions. Here, the swallow frame features were labeled as '0' and concatenated silent and non-swallow frame features as '1'. After applying min-max normalization, the training model was built with the help of binary SVM optimization (449).

In the test case, as it can be understood from Figure 4.14, the features were extracted after framing and windowing operations. Then, using the normalization parameters (minimum and maximum values of features) obtained in the training stage, each feature vector was given as input to the SVM classifier (451) so as to be assigned a class with a posterior probability value. In the end, different merging algorithms to detect boundaries of swallow activity were implemented.

4.2.2 3-class SVM Classifier

Inherently, SVM is a binary classifier, yet multi-class problems can be solved with one-versus-all (OVA) or one-versus-one strategy. Let K be the number of classes, then K binary classifiers are trained in OVA method. In other words, each class has its own classifier in which instances belonging to that class is labeled as positive and the rest as negative. Based on this, silence features were separated from the non-swallow ones and treated as another class, thus, increasing the number of classes from two to three. After the silence features were labeled as '2', each classifier was trained separately with the same optimization technique (449).

The process up to the classifier and thereby the formation of the input were the same

as in binary classification. Unlike, the normalized features were passed through three different classifiers to obtain two score values for each, but only the scores of the corresponding class were taken into account. In the end, classifier outputs of the frames will be three distinct score values (one per class) helping construct different merging algorithms.

Pseudo codes for formation of feature observation matrix, building training model and classifying of each frame in a test signal are provided in Algorithm 7, 8, 9.

Algorithm 7: Creation of Feature Observation Matrix in SVM

input : Collection of audio records K
output: $featMatrix$, a matrix that contains audio features of frames
for $\forall i \in K$ **do**
  $K_i \leftarrow i^{th}$ audio record
  for $\forall f \in K_i$ **do**
    Compute MFCC values of f^{th} frame
    Compute spectral Centroid of f^{th} frame
    Concatenate features
    Add feature row to observation matrix (swallow and non-swallow)
  end
  Extract silent portions of the K_i
  numSilenceFrames = # of silence frames in K_i
  for *k = 1,2 ... numSilenceFrames* **do**
    Compute MFCC values of k^{th} frame
    Compute spectral centroid of k^{th} frame
    Concatenate features
    `// if binary -> silence label is same as non-swallow, o.w. new label`
    Add feature row to observation matrix with its labels (silence)
  end
end
return $featMatrix$

Algorithm 8: Build a classification model

input : Unnormalized observation matrix and labels, $featsMatrix$, # of classes,K, Classifier type,$modelType$
output : Classification model, $svmModel$
parameters: C is a penalty factor for mis-classifying the the training instances,γ controls the effect of a single support vector while drawing the decision boundary. Large (small) gamma implies high (low) bias and low (high) variance
$normObsMatrix \leftarrow$ `MinMaxNormalization(`$obsMatrix$`)`
if $modelType ==$ *'Binary* **then**
  $svmModel \leftarrow$ `svmtrain(`$normObsMatrix$,$labels$, C, γ`)`
else if $modelType ==$ *'one-versus-all (OVA)* **then**
  for $\forall i \in K$ **do**
    $newLabels \leftarrow$ `ThreeToTwoClassLabelConverterOVA(`$labels$`)`
    $svmModel[i] \leftarrow$ `svmtrain(`$NormObsMatrix$,$newLabels$, C_{OVA},γ_{OVA}`)`
  end
Function `MinMaxNormalization(`$obsMatrix$`)`
  $minimums \leftarrow$ Find minimum values of each feature
  $maximums \leftarrow$ Find Maximum values of each feature
  $ranges \leftarrow maximums - minimums$
  $meanExtractedObsMatrix \leftarrow$ subtract each observation of $ObsMatrix$ from minimums
  $newObsMatrix \leftarrow$ divide each observation of $meanextractedObsMatrix$ by ranges
  return $newObsMatrix$

Algorithm 9: Extracting the class scores of each frame in a test signal

input : test audio record, $testSignal$, $SVMModels$,
output: $frameDecisionMatrix$, a matrix containing the classification results of each model of each frame in the test signal

numFrames = # number of frames in test signal
for *k = 1,2 ... numFrames* **do**
    Compute MFCC values of k^{th} frame
    Compute spectral Centroid of k^{th} frame
    Concatenate features
    Add row to test observation matrix
end
$normTestObsMatrix \leftarrow$ `MinMaxNormalization(`$testObservationMatrix$`)`
for *k = 1,2 ... # of models* **do**
    $frameDecisionMatrix[k] \leftarrow$ `svmtest(`$normTestObsMatrix$`,`$SVMModels[k]$`)`
end

4.2.3 Merging Frame Outputs

In the previous subsections of the second classification approach, two different SVM-based implementations are explained. In this part, the determination of swallow boundaries by merging the outputs of the classifiers with different algorithms is explained. In the binary SVM classifier, two different scores ($P(Sw|x)$, $P(NSw|x)$) with a sum equal to 1 were found. In the multi-class, three different scores were obtained separately from the classifiers, yet they were normalized with a soft-max function to make a total sum equal to 1.

When selecting the merge algorithms, how the frame-based classifier behaves in the test data was observed and compared with the ground truth. In Figure 4.15, a portion of a sample recording with three swallow instants and corresponding posterior probabilities for both classifiers are shown.

Taking the results shown in Figure 4.15 into account, different types of merging algorithms were implemented. The closing algorithm, dilation followed by erosion, was used with minor modifications. Furthermore, median and moving average filters were applied in some cases. Moreover, dedicated finite state machine algorithms which are inspired by the previous generic algorithms were used to increase boundary detection performance.

Closing Operation (Dilation + Erosion): This operation is usually applied to binary images to enlarge the area of the bright regions, yet it can also be performed to one-dimensional signal. For that purpose, two inputs, the signal (binary image or sequence) and a kernel which determines the nature of operation are required. And,

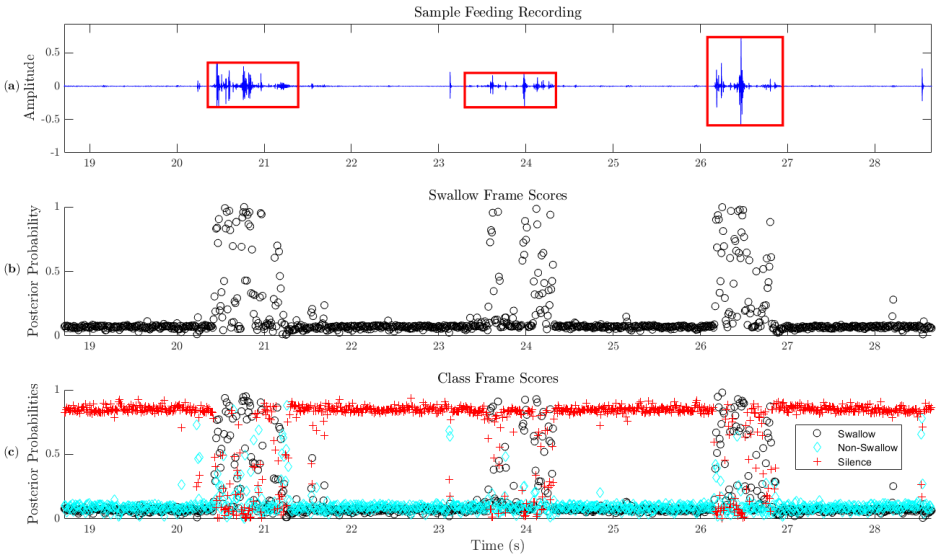


Figure 4.15: (a) A section of feeding signal including three swallow activities. The red colored rectangles are the ground truth of swallow boundaries. (b) Swallow class posterior probabilities from Binary SVM Classifier. (c) Posterior probabilities of each class obtained from three-class SVM.

dilation and erosion, are the two basic morphological operators constructing this operation. Dilation of a signal, x, by a sliding kernel, $k = [1, 1, \ldots, 1]$, generates a new signal in which the zeros are transformed to ones according to rules given [60].

1. If the center of the kernel coincides with a sample with value 0, no change; move to next sample.
2. If the center of the kernel coincides with a sample with value 1, change all samples covered by kernel as 1.

The erosion operation is similar to dilation yet the action taken due to the second condition is different. The rules of erosion are given below.

1. If the center of the kernel coincides with a sample with value 0, no change; move to next sample.
2. If the center of the kernel coincides with a sample with value 1 and at least one kernel element intersects with a zero sample, change all samples covered by kernel as 0.

In this study, two class probability estimates were converted into binary outputs by thresholding method. After that, dilation and erosion operations were applied with a kernel vector of all ones to fill holes.

Moving Average Filter: In general, the moving average (MA) filter is used to eliminate noisy data or smooth out the input data [61]. Since it is practical and many times successful, it is one of the most commonly used filters in the Digital Signal Processing area. The mathematical form of the MA filter can be written in the form of the following difference equation

$$y[n] = \frac{1}{L} \sum_{k=0}^{L-1} x[n-k] \tag{456}$$

where $x[n]$ is the n^{th} sample of the input signal x and y corresponds the output signal.

In this study, the frame scores were filtered with MA filter and output values were converted to binary outputs according to a certain threshold. After that, consecutive ones were merged to construct the boundaries of the swallow activities.

Median Filter: In the moving average filter, the average of the previous samples and the current input forms the output samples [62]. Remember that the median is the middle value separating the greater and lesser halves. For L is an odd integer, the n^{th} sample of y is calculated as

$$y[n] = x\left[n - \frac{L+1}{2}\right]. \tag{457}$$

Thresholding was first applied to the score vector, unlike the MA filter. After that, nonlinear-digital median filtering was applied to the binary predictions. Similar to the MA case, consecutive ones were merged to form boundaries.

2-class Finite State Machine The closing algorithm, moving average and median filters are mostly used for generic cases and may give a lot of false alarms due to the existence of special scenarios during the feeding. To deal with, a finite state machine (FSM) algorithm for two classes was constructed. The algorithm 10 shows the pseudo-code for two class FSM.

The input is a binary vector obtained by thresholding the probability estimate values

Algorithm 10: FSM algorithm with 2 classes to extract swallow boundaries

```
   input  : Binary Frame Prediction Vector, fdv
   output : Boundary Matrix, bm
1  Initialize Counters
2  Specify Thresholds
3  for ∀i ∈ range(fdv) do
4      prediction = fdv[i]
5      switch state do
6          case 0 do
7              if prediction == 1 then
8                  No_1 = No_1 + 1
9                  Suc_1 = Suc_1 + 1
10                 start = i
11                 state = 1;
12         end
13         case 1 do
14             if prediction == 0 then
15                 end = i − 1
16                 no_0 = no_0 + 1
17                 suc_1 = 0
18                 state = 2
19             else
20                 no_1 = no_1 + 1
21                 suc_1 = suc_1 + 1
22             end
23         end
24         case 2 do
25             if prediction == 0 then
26                 if no_0 < no_0TH then
27                     no_0 = no_0 + 1
28                     state = 2
29                 else
30                     state = 0
31                     if no_1 > no_1TH and suc_1 > suc_1TH and No_1/No_0 > oneZeroTH then
32                         bm.append(start, finish)
33                     Reset Counters
34                 end
35             else
36                 state = 1
37                 no_1 = no_1 + 1
38                 no_0 = 0
39                 suc_1 = 1
40             end
41         end
42     end
43 end
```

of each frame. As can be understood from the pseudo-code, the number of ones and consecutive ones, the number of zeros were taken into consideration to determine the boundaries of swallow action.

3-class Finite State Machine: As it can be observed from the figure 4.15, a probability estimate value was extracted for each class for a frame. For that reason, a dedicated FSM algorithm for three class was designed. The algorithm 11 shows the

pseudo-code for three class FSM.

Algorithm 11: FSM algorithm with 3 classes to extract swallow boundaries

```
input  : Frame Prediction Scores , fps(N × 3)
output : Boundary Matrix, bm
Initialize Counters
Specify Thresholds
for ∀i ∈ range(fps) do
    scoreVec = fps[i] // 1st -> Sw, 2nd -> Non − SW, 3rd -> Sil
    // Conditions
    c1 = scoreVec[1] > swTH
    c2 = scoreVec[1] > nswTH
    c3 = scoreVec[1] > silTH
    switch state do
        case 0 do
            if c3 and c1 then
                state = 1
                start = i
                swCnt = swCnt + 1
        end
        case 1 do
            if c3 and (c1 or c2) then
                state = 1
                end = i
                if c1 then
                    swCnt = swCnt + 1
                else
                    nswCnt = nswCnt + 1
                end
            else
                state = 2
            end
        end
        case 2 do
            if c3 and (c1 or c2) then
                state = 1
                end = i
                idleCnt = 0
                if c1 then
                    swCnt = swCnt + 1
                else
                    nswCnt = nswCnt + 1
                end
            else
                idleCnt = idleCnt + 1
                if idleCnt > idleTH then
                    if swCnt > nswCnt − swNswTH and swCnt + nswCnt > sumTH then
                        bm.append(start, end)
                    Reset Counters
            end
        end
    end
end
```

Let N be frame number of an acoustic signal obtained from a feeding session, then the input for the FSM algorithm will be $N \times 3$ matrix including frame probability values

for each class. Swallow and non-swallow frame counts and thresholds for each class were taken into consideration to construct this algorithm.

CHAPTER 5

EXPERIMENTS AND RESULTS

In this chapter, the results of the segmentation, feature extraction and classification tasks are given. To detect swallow activities from the feeding signal, two different approaches are described in the previous chapter. The experimental procedures for both swallow detection mechanism are shown in Figure 5.1 in a block diagram. In this chapter, the implementation of algorithms in each block is explained and the results are given both statistically and visually.

MATLAB software was utilized for each module given in 5.1 except for training the SVM models [63]. A multi-threaded function for 5-fold cross validation was implemented in C++ platform with the help of LIBSVM tool [58]. Therefore, the higher the number of CPU cores, the greater the speed of cross-validation algorithm.

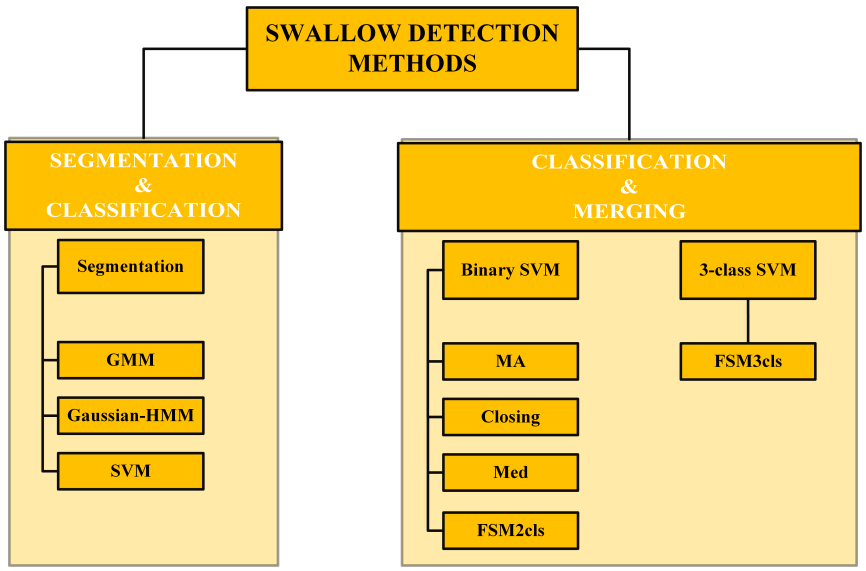


Figure 5.1: Experimental procedures for both swallow detection mechanisms

5.1 Segmentation

In this section, the calculation method of both precision (positive predictive value) and the recall (true positive rate) values are explained, then the precision-recall scatter plot of both segmentation subsystems mentioned in 4.1.1.1 and 4.1.1.2 are plotted with different parameter combinations. After finding the optimum parameter combinations, detected sound activities are visually depicted.

5.1.1 Evaluation metrics

As it is mentioned in the segmentation section of the methodology chapter, several parameters are utilized for both segmentation systems. The first one has three different threshold parameters whereas the finite state machine algorithm of the second one requires four different parameters. For that reason, it is necessary to extract parameter combinations in which the best performance is achieved.

Detecting the boundary of swallow events of adults is available in the existing studies and while calculating the performance, the majority of them utilize correct or incorrect swallow event numbers. Some also measure the performance considering the localization error such as [38, 39] yet their evaluation metrics are method-specific. On the other hand, as far as we know, there is no available segmentation system paying attention to localization issues for the infants and no evaluation metrics regarding the boundaries are found. However, since this problem is an audio segmentation problem, the evaluation metrics of audio segmentation as in [64, 65] were applied.

Precision-Recall can be considered as a useful measure of prediction for this problem. The popular definitions of precision, P, and recall, R are given as

$$P = \frac{TP}{TP + FN} \qquad\qquad R = \frac{TP}{TP + FP} \tag{51}$$

where TP is the number of true positives, FN and FP are the number of false negatives and false positives respectively.

However, in this study, these three parameters were not considered as numbers but time duration as depicted in Figure 5.2. Furthermore, the precision and recall for

each segment can be calculated as

$$P = \frac{\text{Detected} \cap \text{Ground truth}}{\text{Detected}} \qquad R = \frac{\text{Detected} \cap \text{Ground Truth}}{\text{Ground truth}} \tag{52}$$

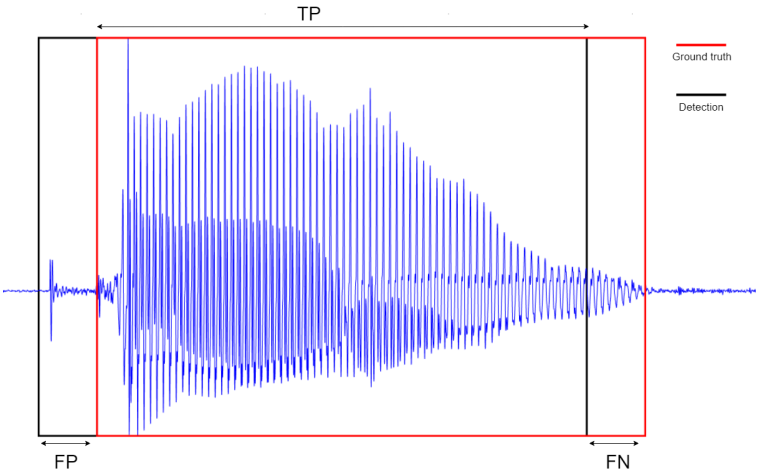


Figure 5.2: Illustration of TP, FP and FN on an example of non-swallow sound activity in terms of time duration

In order to calculate recall for a specific parameter combination, the intersection of each labeled sound activity with the sound intervals obtained as a result of the segmentation algorithm was computed. In other words, for each ground truth interval, a recall value was estimated and the average of them determined the recall value of the corresponding parameter combination. On the other hand, precision value was computed by focusing on the detected intervals. For each sound segment, a precision value was estimated by looking at the intersection of ground truth sound intervals and similarly, the mean of them specified the precision value of the corresponding parameter combination.

5.1.2 Energy Based Segmentation Algorithm:

The frame length was selected to be 200 ms and the time duration between two successive frames was 10 ms. In this algorithm, detection performance was affected by four different parameters given as

1. **Min Peak Distance** (mpd)**:** Distance between two successive peaks cannot be smaller than mpd,

2. **Threshold 1** ($TH1$)**:** Threshold to ignore small energy peaks,

3. **Threshold 2** ($TH2$)**:** Related to width of the detected boundaries,

4. **Threshold 3** ($TH3$)**:** For precise adjustment of the detected boundaries.

All parameters were swept together through a range of values and the Precision-Recall scatter plot was obtained as shown in Figure 5.3 so as to find the best parameter combination in terms of F1 Score defined by

$$F1 = \frac{2PR}{P+R}. \tag{53}$$

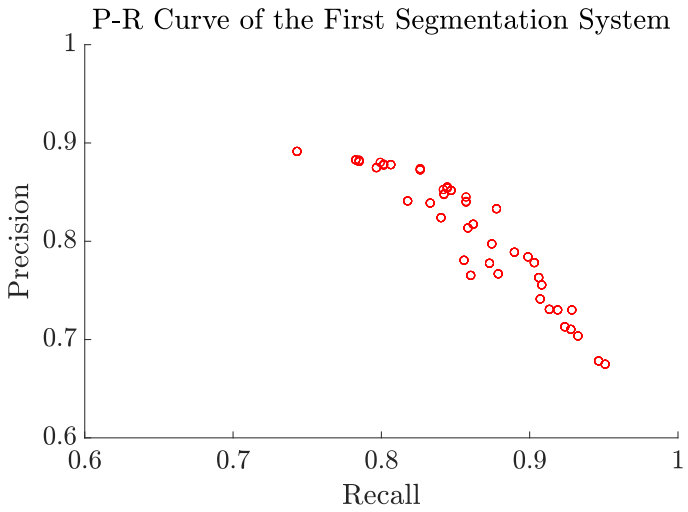


Figure 5.3: P-R scatter plot of the Energy Based Segmentation System

The R and P values of satisfying the best performance were found to be 0.87 and 0.83 respectively. In Figure 5.4, two sample recordings, which were not utilized in the parameter sweep procedure, and black colored segmented intervals are shown. The results were obtained using the optimum parameter combination, giving the best precision and recall values.

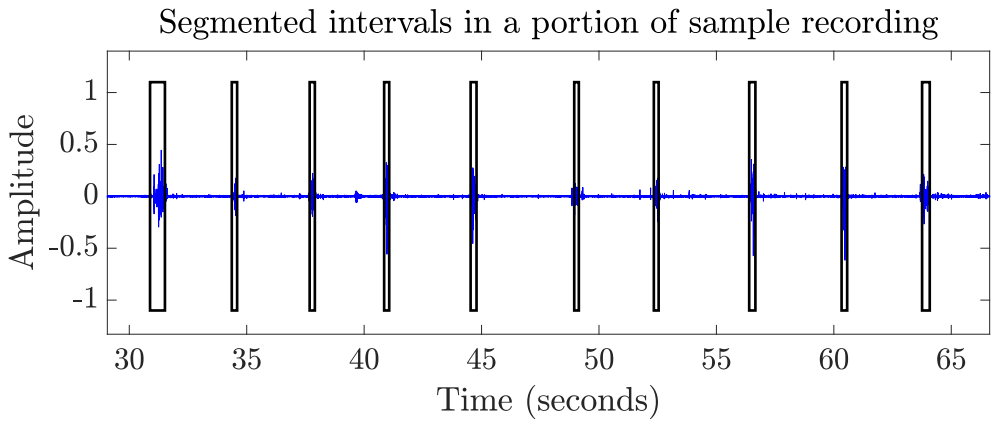


Figure 5.4: An example of successful segmentation in a portion of feeding signal including both swallow and non-swallow activities. The black colored rectangles represents the detected boundaries of the Energy Based Segmentation System.

5.1.3 Pattern Recognition Based Segmentation Algorithm

Audio recordings of 10 infants were utilized for both train and test parts of this algorithm. Both the frame length and time duration between two successive frames were selected to be as 10 ms (no overlap) as stated in [42]. In order to merge binary outputs of the classifier to finish the segmentation process, the finite state machine algorithm was used. The performance of this detection subsystem was also affected by various parameters shown in Figure 4.7.

Similarly, the parameter sweep procedure was applied and the Precision-Recall curve of the system obtained as depicted in Figure 5.5.

The R and P values of satisfying the best performance were found to be 0.70 and 0.93 respectively. In Figure 5.6, two different portions from two feeding recordings apart from the train data set, were shown. As in the energy based one, black colored rectangles are the boundaries found by pattern recognition based segmentation algorithm with optimum parameter combination.

Based on the Precision-Recall scatter plot of both segmentation subsystems and several visual inspections, both algorithms may have advantages or disadvantages for different purposes. On the other hand, it is observed that recall values of the feeding recordings of which are not in the train data set of the latter system were lower than

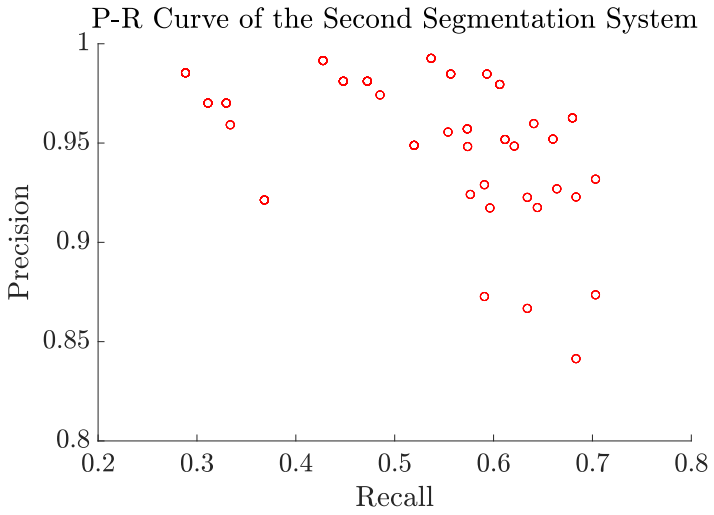


Figure 5.5: P-R Scatter Plot of the Pattern Recognition Based Segmentation System

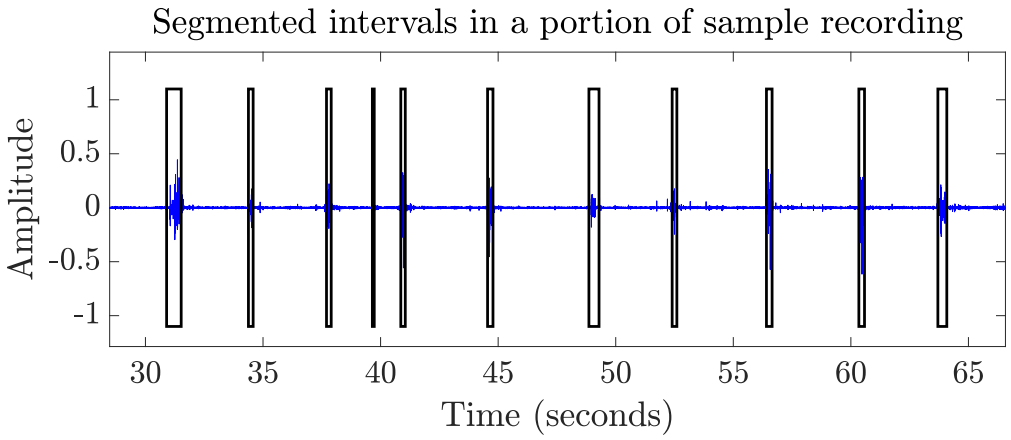


Figure 5.6: An example of successful segmentation in a portion of feeding signal including both swallow and non-swallow activities. The black colored rectangles represents the detected boundaries of the Pattern Recognition Based Segmentation System

expected. Moreover, F1 score of the former one is slightly higher. Hence, the first segmentation algorithm was used as a pre-processing stage before feature extraction block.

5.2 Segmentation followed by Classification Experiments

The proposed classification techniques require training models to classify frames of the segmented intervals. Hence, each classifier was trained from a data set consisting of 35 recordings of the term infants. The total number of labeled intervals in train data set is 1221, 642 of which are swallow activities and 579 of are non-swallow ones. Furthermore, each interval was divided into overlapping frames of duration 25 ms and 10 ms stride, thus leading the total numbers of frames belonging to swallow and non-swallow classes are 15704 and 14584 respectively.

5.2.1 Feature Selection and EM-GMM

This section explains how to select features or to eliminate redundant ones from the observation matrix. For this purpose, a small portion of the above data set was used for both train and test processes with the help of EM-GMM algorithm. Classification models were trained and tested wheh $K = 4$ for both classes. The list of four combinations to choose the best performance in a mini data set is given as

1. MFCC, Spectral Centroid, Spectral Spread, Spectral Flatness
2. MFCC, Spectral Centroid, Spectral Spread
3. MFCC, Spectral Centroid
4. Spectral Centroid, Spectral Spread.

In the mini train data set, there are 200 labeled intervals belonging to swallow class and 160 for non-swallow, while in test data set these numbers are 100 and 80. After the selection of intervals, the observation matrices for each class were constructed and the mean vectors, covariance matrices and initial probability estimates were estimated. In the end, each interval in the test data set was assigned to a class according to decision rules, the sum of log-likelihood and majority voting. The accuracy of each combination was calculated according to the correct classification rate for both classes. The number of true and false classification numbers are given in the 5.1.

Table 5.1: Classification results of mini data set on EM-GMM

Feature Combination	True		False	
	Sum of log-liks	Majority Voting	Sum of log-liks	Majority Voting
MFCC, Spectral Centroid, Spectral Spread, Spectral Flatness	145	124	35	56
MFCC, Spectral Centroid, Spectral Spread	161	161	19	19
MFCC, Spectral Centroid	**170**	**166**	**10**	**14**
Spectral Centroid, Spectral Spread	132	119	48	61

After the optimum feature combination was selected as 14-dimensional MFCC-Spectral Centroid combination, both segmentation followed by classification and classification followed by merging approaches used them as inputs.

5.2.2 Selection of the Best Classifiers

EM-GMM: For a 14-D observation matrix of swallow and non-swallow classes, distribution of clusters cannot be visualized and there is no prior information related the number of Gaussian mixtures, K. Hence, multiple EM-GMMs were trained with different K values and those models were used to make predictions on data not used in the training procedure by following the 5-fold cross-validation approach in order to prevent overfitting. The decision rule type while testing the model was another parameter affecting the performance as well.

Note that, cross-validation procedure was applied on labeled intervals. After the intervals were separated, the frame feature vectors of all intervals belonging the same fold were merged for train phase yet, not the frames, intervals were tested to achieve the best performance.

Gaussian-HMM: Remember that the existence of the state transition probability matrix is the essential difference between EM-GMM and Gaussian-HMM for this study. Therefore, for N state, continuous normal density parameters were estimated. Again, 5-fold cross-validation technique was applied and the N parameter was swept to train multiple Gaussian HMMs for both classes. On the other hand, the delta variable added to the output likelihood of the Gaussian HMM of swallow class was tested

with different values. Since HMMs needs temporal information, intervals were used to both train and test sessions.

SVM: To map the original data set into a higher dimensional space, Radial Basis Function (RBF) kernel, which is popular for kernelized learning algorithms, was used. The γ parameter of RBF kernel can be considered as a similarity measure between two different points. Small γ increases variance of Gaussian function yielding high similarity between two points, although they are far from each other. Varying the γ parameter together with the penalty factor, C, of SVM, 5-fold cross-validation procedure was applied and for each pair, an SVM model was trained.

As in the case of GMM, the approach was "frames for the train" and "intervals for test".

To compare performances of training models for each classifier, precision and recall values were found according to 51. And, the calculation method of inside parameters is given below.

1. TP: Number of swallow intervals classified as swallow
2. FP: Number of non-swallow intervals classified as swallow
3. FN: Number of swallow intervals classified as non-swallow

Cross validation procedure was applied to data set such that frames belonging to the same interval stayed in the same fold. The Precision-Recall scatter plots of all classifiers with 5-fold cross-validation on labeled intervals are depicted in Figure 5.7.

Maximum F1 scores and corresponding recall, precision values are shown in Table 5.2. Also, the parameter combinations for the best classifiers having the maximum F1 score can be seen in the following list.

1. GMM: $K = 4$, decision type = sum of log-likelihoods
2. Gaussian-HMM: $K = 4$, $\delta = 0.5$
3. SVM: $C = 2^7$, $\gamma = 2^2$

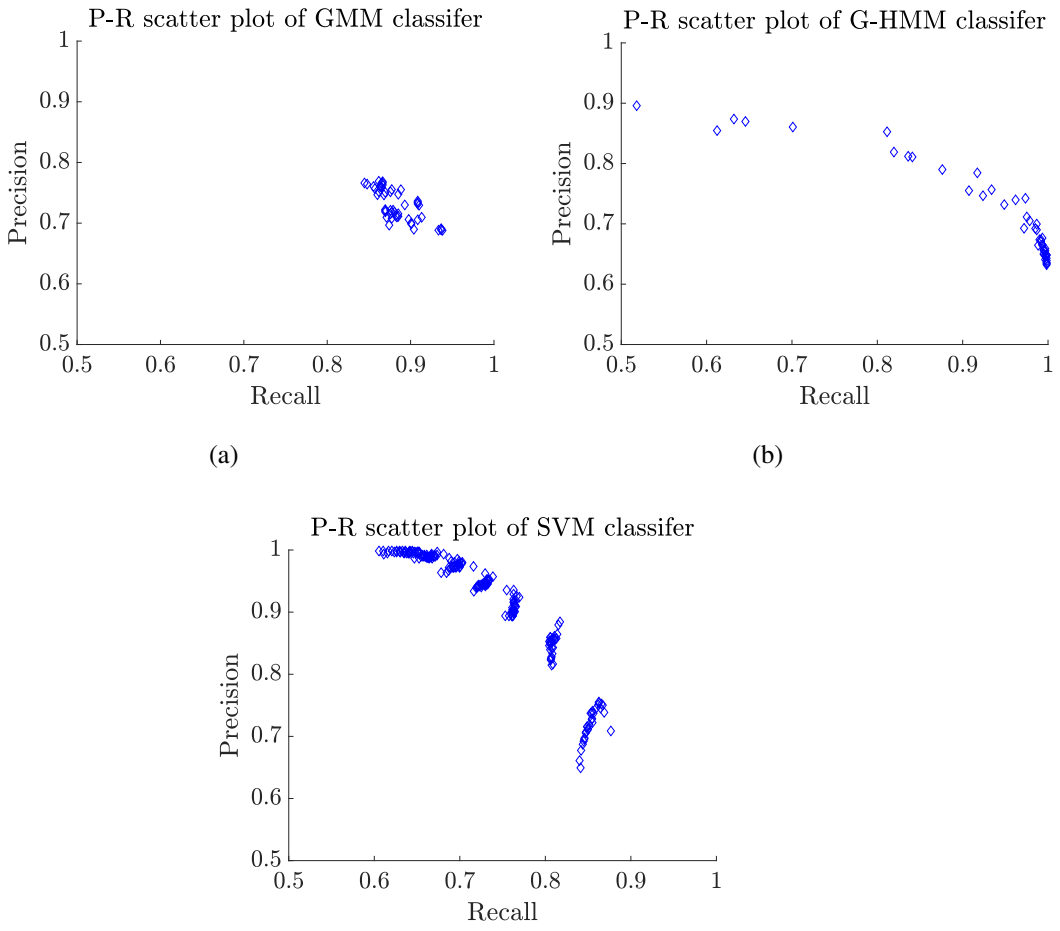


Figure 5.7: Precision-Recall scatter plot of all classifiers from 5-fold cross validation. (a) EM-GMM, (b) Gaussian-HMM, (c) SVM

Table 5.2: The best performance metric values of three classifiers after parameter optimization

	GMM	**Gaussian-HMM**	**SVM**
Recall	0.888	0.917	0.817
Precision	0.755	0.784	0.884
F1 Score	0.816	0.846	**0.849**

5.2.3 Assessment of the Best Classifiers with Segmentation

The cross-validation method only deals with manually labeled intervals and selects the best models for each learning algorithm. On the other hand, remember that the primary purpose is to detect swallow intervals. Thus, segmentation and classification blocks were put together to detect only swallow activities.

The parameters giving the maximum F1 score were used to segment sound activity intervals at first. In other words, the pre-processing stage for each classifier was identical. Furthermore, for each classifier, the class of the segmented interval predictions was made using the corresponding optimum parameters.

Performance assessment of three classifiers together with the segmentation algorithm is similar to the approach given in Figure 5.2. After the class of the segment is determined, the non-swallow ones are eliminated and swallow ones remains. Then, the precision and recall values are calculated for each classifier with their optimum parameters and shown in Table 5.3.

Table 5.3: Performance values classification followed by segmentation approaches

	Seg+GMM	**Seg+Gaussian-HMM**	**Seg+SVM**
Recall	0.588	0.603	0.643
Precision	0.616	0.691	0.667
F1 Score	0.601	0.644	**0.655**

5.3 Classification followed by Merging Algorithms

Remember that the proposed techniques in this approach do not have any pre-processing mechanism like segmentation. Firstly, the entire feeding signal was split into overlapping frames of 25 ms with a stride of 10 ms. Then, the features (MFCC, Spectral Centroid) of each frame were extracted. After the classifiers (Binary SVM and 3-class SVM Classifier) assigned each instance to a class, the merging algorithms given

below were applied to detect swallow boundaries.

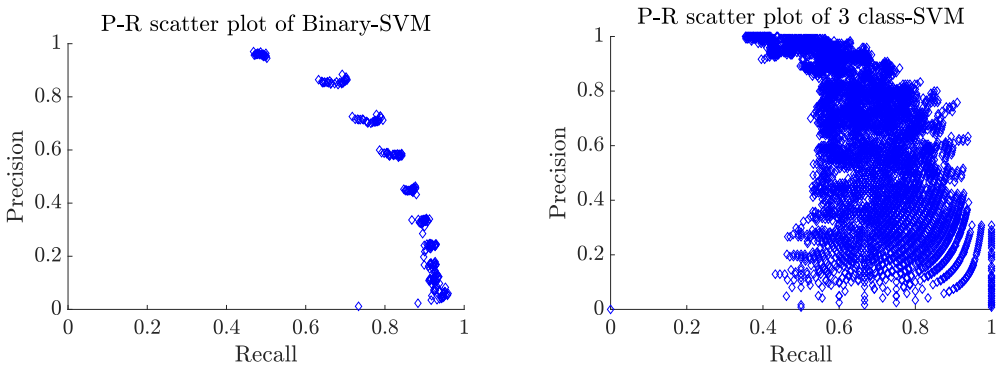
1. For Binary SVM Classifier
    - Closing Operation (Dilation and Erosion)
    - Moving Average Filter
    - Median Filter
    - Finite State Machine algorithm for 2 classes
2. For 3-class SVM Classifier
    - Finite State Machine algorithm for 3 classes

5.3.1 Selection of the Best Classifiers

Binary SVM: To select the best classifier, RBF kernel was utilized and 5-fold cross-validation technique was applied with different C-γ pairs as described in the previous section. However, silence intervals were also labeled as non-swallow instants. Combining all the frames of intervals allocated for training and test intervals for each fold, the Precision-Recall scatter plot was extracted.

3-class SVM: Unlike the binary SVM approach, silence intervals were considered to be another class. Since the approach for multi-classes is "one-versus-all" (OVA), three different models (one per class) were trained for each C-γ pair. In the test case, two different δ values (δ_1,δ_2) were swept to be used when comparing the logarithm sum of probability estimates. Since three models are required to make a prediction, optimum C-γ pairs were extracted separately for each class.

Recall and precision values were computed and Precision-Recall scatter plots of both classifiers with 5-fold cross-validation on labeled intervals are depicted in Figure 5.8. Also, 5.3 shows the precision and recall values giving the best performance (maximum F1 score) after parameter optimization.



(a) (b)

Figure 5.8: Precision-Recall scatter plot of both classifiers from 5-fold cross validation. (a) Binary-SVM, (b) 3-class SVM

Table 5.4: The best metric values of two classifiers after parameter optimization

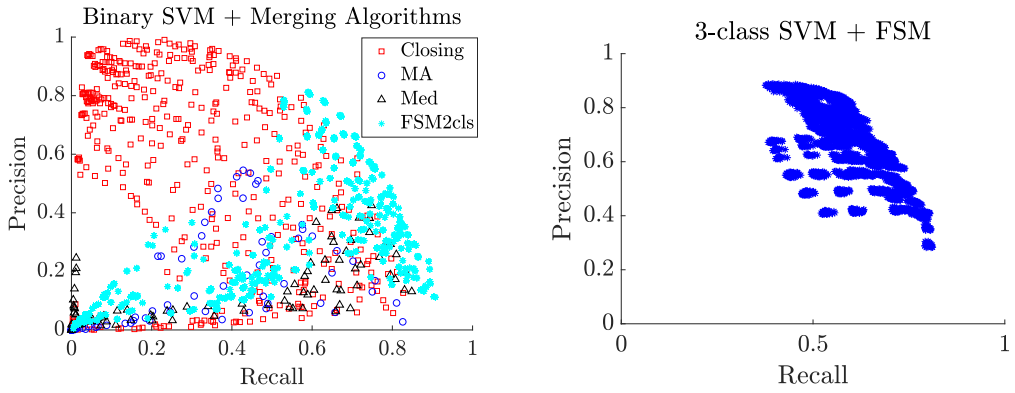
	Binary-SVM	**3-class SVM**
Recall	0.702	0.857
Precision	0.874	0.8
F1 Score	0.7786	**0.826**

5.3.2 Assessment of the Best Classifiers with Merging Algorithms

The main goal of the previous section is to select the best training models by tuning the parameters affecting classification performance. However, it is assumed that the intervals of each class were already segmented. In Figure 4.15, three swallow instants are shown in a portion of sample recording. Furthermore, frame probability estimates for swallow class in binary SVM for three classes in 3-class SVM are depicted in that figure.

To achieve the main goal, probability estimate values were merged with given algorithms. Evaluation of the performance is the same with the segmentation followed by

classification approach and the Precision-Recall scatter plots together with the merging algorithms are displayed in Figure 5.9.



(a) (b)

Figure 5.9: (a) P-R scatter plot of the binary SVM classifier + merging algorithms, (b) 3 class SVM + Merging algorithm (FSM3cls)

It is evident that moving average and the median filters fail to combine frame probability estimate values of binary SVM classifier compared to closing and FSM algorithms. Also, FSM algorithm appears to be slightly more successful than closing (dilation+erosion).

In summary, to find the swallowing instants from the feeding signal of an infant, a total of 8 different paths were followed, three from segmentation followed by the classification, five from classification followed by merging. The comparison of the time duration based performance results of all paths is given in Table 5.5. In addition, event based evaluation metrics were also computed for the pipelines of both approaches as given 5.6. In this case, it does not matter how precise a swallow event is detected.

Outputs of the swallow boundary detection for two approaches were also observed. In Figure 5.10 and 5.13, cases where the detection systems performed successfully can be seen and the segments determined by the algorithms are shown in pink rectangles while the ground truth is green. In Figure 5.11 and 5.14, false positive examples are displayed. Another scenario is that the detected segment contains two swallow moments, as indicated in Figure 5.12. On the other hand, three swallow events are

Table 5.5: Time duration based swallow boundary detection performance of eight different paths

	Segmentation followed by Classification			**Classification followed by Merging**				
	Seg+GMM	**Seg+Gaussian-HMM**	**Seg+SVM**	**B-SVM+Closing**	**B-SVM+MA**	**B-SVM+Med**	**B-SVM+FSM2Cls**	**3-SVM+FSM3cls**
Recall	0.588	0.603	0.643	0.713	0.446	0.748	0.6703	0.653
Precision	0.616	0.691	0.667	0.586	0.535	0.425	0.733	0.752
F1 Score	0.601	0.644	0.655	0.643	0.487	0.542	**0.700**	0.6981

Table 5.6: Swallow event detection performance of eight different paths

	Segmentation followed by Classification			**Classification followed by Merging**				
	Seg+GMM	**Seg+Gaussian-HMM**	**Seg+SVM**	**B-SVM+Closing**	**B-SVM+MA**	**B-SVM+Med**	**B-SVM+FSM2Cls**	**3-SVM+FSM3cls**
Recall	0.682	0.714	0.74	0.808	0.615	0.642	0.862	0.87
Precision	0.783	0.738	0.827	0.793	0.652	0.598	0.763	0.77
F1 Score	0.729	0.726	0.781	0.80	0.633	0.614	0.81	**0.817**

missed by the detection system for the case given in Figure 5.15.

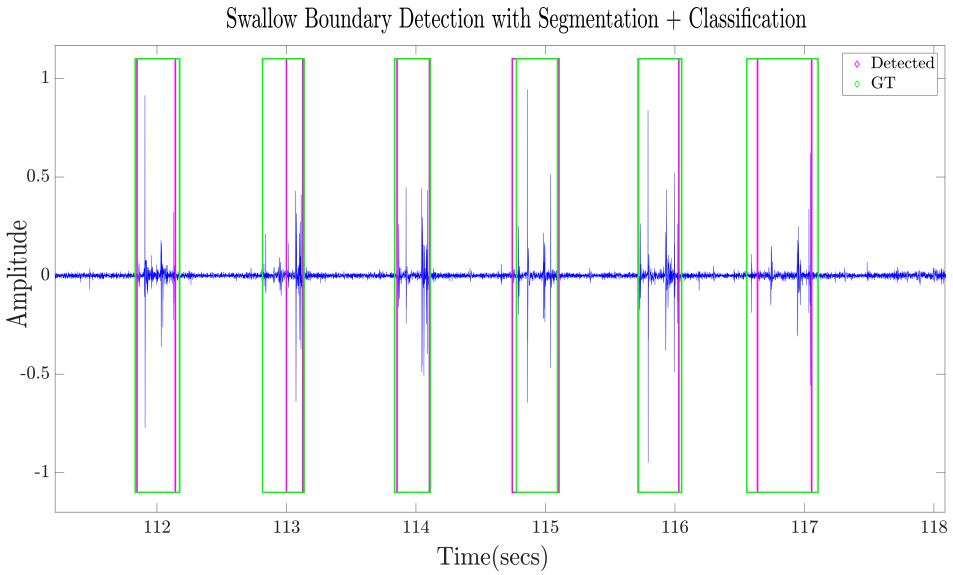


Figure 5.10: Sample feeding signal including 6 swallow events. Green rectangles represent manually labeled ground truth intervals, pink rectangles are for segmented swallow events

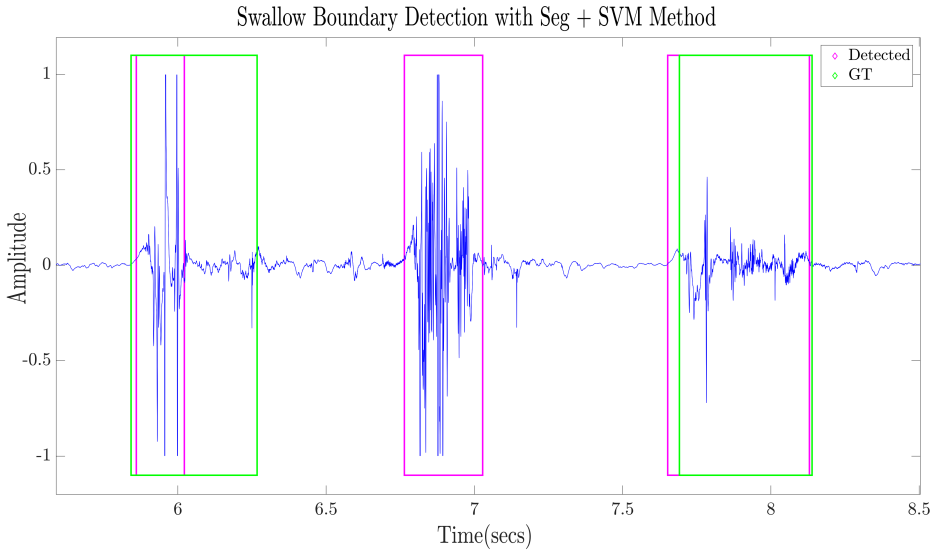


Figure 5.11: Segmentation + Classification, a false positive at the middle of three swallow events.

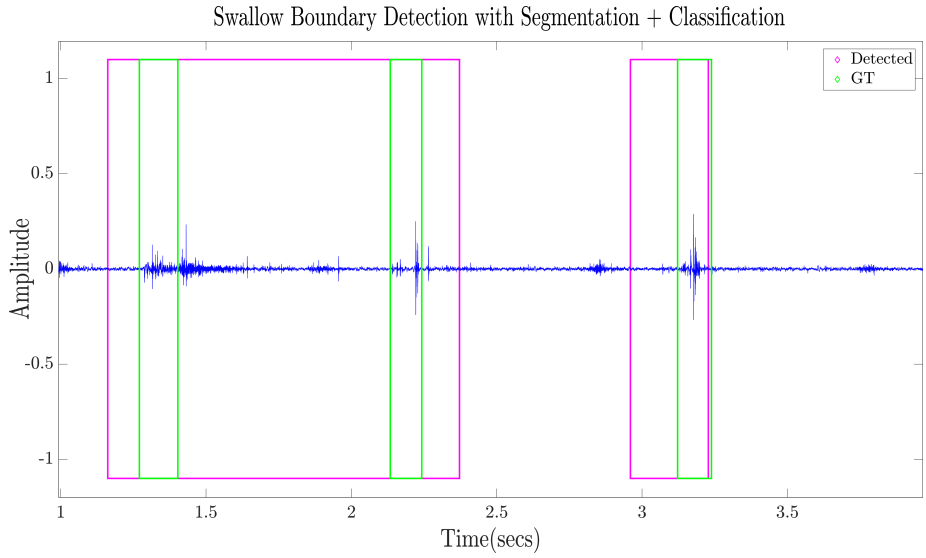


Figure 5.12: Segmentation + Classification, classification failure due to wrong sound activity segmentation

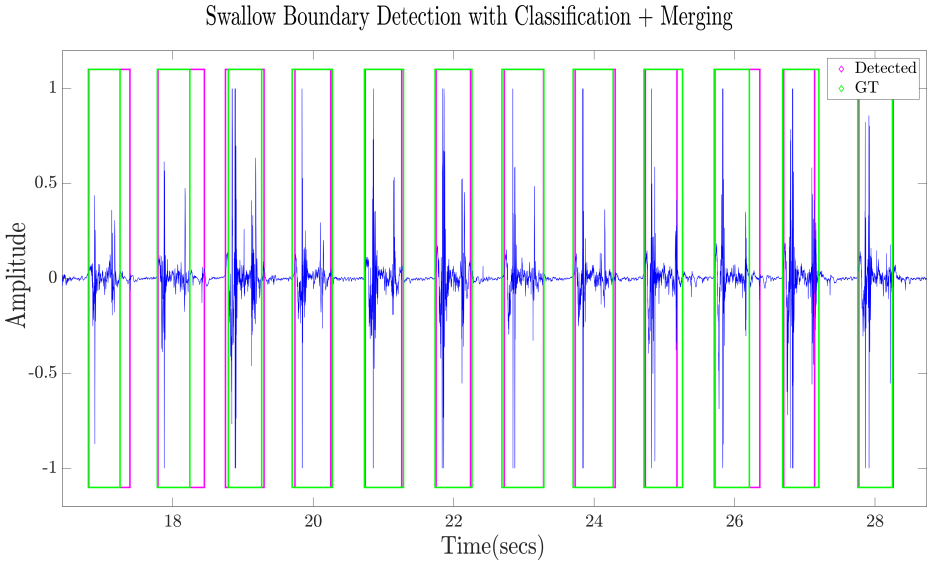


Figure 5.13: Classification + Merging, 12 swallow events which are classified correctly

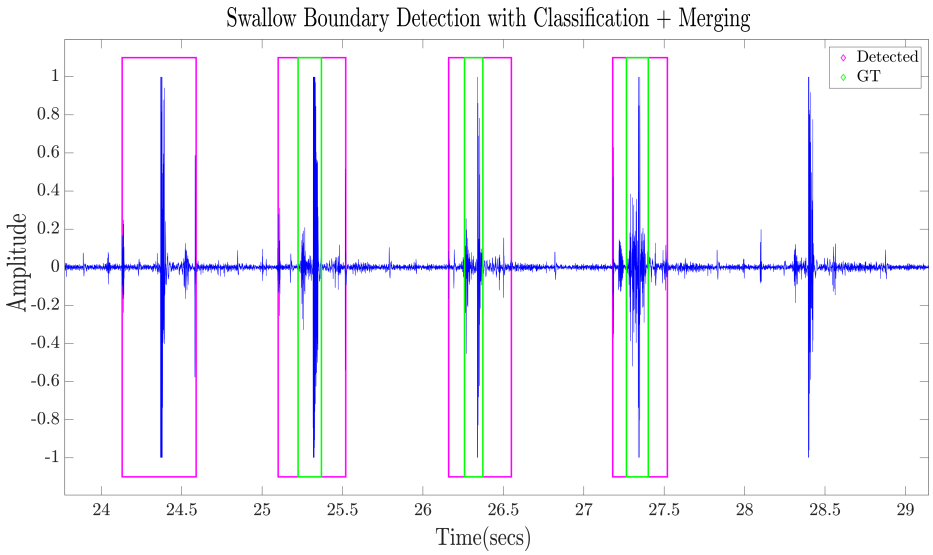


Figure 5.14: Classification + Merging, a false positive at the beginning

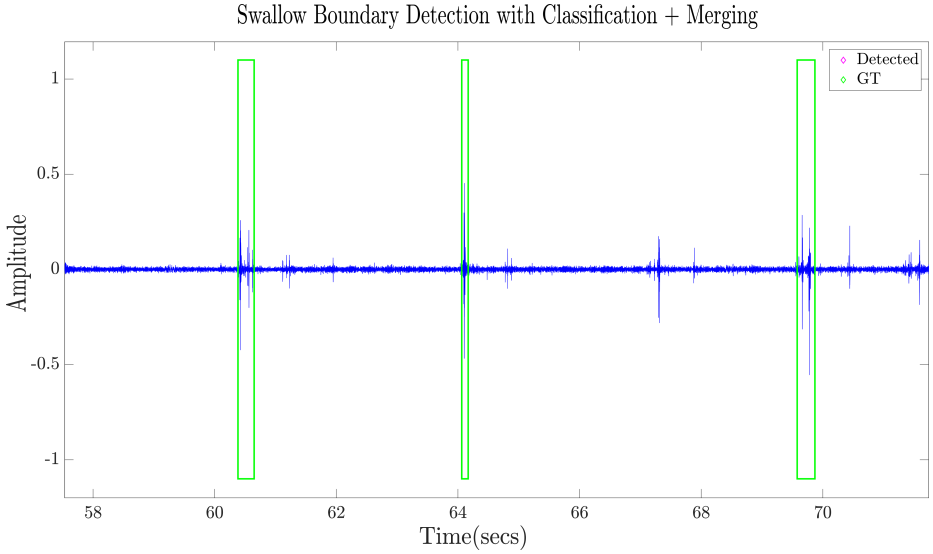


Figure 5.15: Classification + Merging, 3 swallow events that are missed

5.3.3 Discussion

For the first approach, remember that the segmentation part before classification is identical for each method. Hence, it can be inferred that the evaluation metric values, which are given Table 5.5, reflect the classification power of given learning algorithms. On the other hand, the precision-recall values of GMM and Gaussian-HMM support the assumption that the frames of the swallow audio segment are not independent of each other. In other words, the presence of the state transition matrix (Gaussian-HMM) increased the classification performance. However, the discrimination power of the SVM method appears to be slightly better than Gaussian-HMM.

When the sound activities are too close to each other or the feeding signal contains noisy data, the segmentation algorithm may not work properly and the segmentation error can also influence the classifier performance as indicated in Figure 5.12. This may be the reason why the classification followed by merging approach gives more accurate results compared to the former one. On the other hand, ignoring the short-time swallows can be considered as the handicap of the second approach as shown in Figure 5.15

As shown in Figure 3.2b, swallow activities may contain two stages and a short silence period in between. In those cases, corresponding silence frames are not assigned to swallow class, thus decreasing the ability to merge frames by moving average and median filters. Because these filters are expected to work properly when the swallow sounds are in a compact form. Otherwise, it is inevitable to observe more than one detected swallow episode in a swallow episode. Due to the significant increase in the performance, the closing algorithm can be seen as a solution which is prone to such problems. Nevertheless, purpose-specific finite state machine algorithms are described to merge frame outputs and performed more robust compared to all.

Interpreting the silence frames as another third class is thought to perform better compared to two classes case in which silence ones are treated as a non-swallow class. However, as indicated in Table 5.5, their performances are almost equal.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Correct assessment of infants' oral feeding skills to prevent developmental, cognitive problems in the future and minimize the adverse effects of prematurity is of great importance. Swallowing is one of the fundamental phases of the entire feeding process and has an essential role in the evaluation of oral feeding readiness.

Since the maximum number of rhythmic swallows, swallow frequency and the average time between swallows are found to be positively/negatively correlated with feeding maturity, detecting swallow segments automatically from an acoustical feeding signal creates an infrastructure for obtaining the corresponding swallow-related statistics. Hence, this may help neonatal doctors to associate oral feeding readiness of the infants. In this study, two pattern recognition pipelines, segmentation followed by classification and classification followed by merging, are proposed to segment and classify swallow activities automatically.

To differentiate swallow and non-swallow sounds, spectral characteristics are examined. Among the four different features, MFCC and spectral centroid (14-D feature vector) are selected by comparing the performances on GMM classifier for different feature combinations with a mini data set.

In the previous method, sound activities are segmented before classification. For this purpose, two different methods, energy and pattern recognition based, are used. Sound activity detection performances of both are optimized by parameter tuning. The energy-based boundary detection is selected to be optimum in terms of F1 measure. Performance evaluation metrics are calculated by taking the methods used in literature for audio segment problems into consideration. ll

Three classifiers which are GMM, Gaussian-HMM and SVM are applied to segmented intervals to assign them to swallow or non-swallow class. To prevent memorization and overfitting problems for each classifier, 5-fold cross-validation procedure together with parameter sweeping is applied to build train models and test manually labeled intervals.

In the latter method, overlapping small frames are classified before merging according to a rule. For the classification part, two classifiers, binary SVM and 3-class are used to assign each frame to a class. The criteria for selection of best classifiers are the same as ones in the first approach. Then various merging algorithms are utilized to combine the outputs of frames to segment swallow boundaries. After the parameters of merging algorithms are optimized, pipelines are implemented.

The detailed experimental comparison for eight different pipelines is given in Section 5.3.3. It is deduced that the error in the segmentation of the sound activities due to environmental noises or the very short period between activities affects the classification performance adversely. Hence, using the time duration based evaluation metrics, the best cases are obtained in the "binary SVM + FSM" and "3 class SVM + FSM" where F1 measures are almost equal to 0.70. Similarly, same segmentation and classification pipelines provides the best F1 scores (nearly 0.81) in the event based performance evaluation.

Another significant issue is the time complexity of the pipelines. The first approach with the energy-based segmentation can be favored since it only applies classification algorithms on the segmented intervals. However, when the boundaries are detected with the pattern recognition based algorithm, the complexity is slightly higher. On the other hand, when the low time complexity is desired, classification followed by merging algorithms should not be preferred due to the multiplication of relatively high numbers of support vectors with all frame feature vectors.

Although it depends on the implementation platform, all the pipelines other than ones using the energy-based segmentation as a pre-processing stage can be implemented, so that swallow sounds are segmented in real time.

While this thesis presents a comprehensive set of machine-learning based solutions to

detect swallow segments from the acoustic feeding signal and forms a decision support skeleton for neonatal doctors, the study is highly open to further improvements which are listed as follows.

- Representation of swallow sound signal is of great importance to reduce noise effect and enhance the segmentation and classification performances.
- More clinical trials are required to achieve higher performance, thus constructing a more robust infrastructure to obtain swallow-related statistical data.
- Working together with neonatal physicians, a feeding maturity rating system can be implemented thanks to statistical information obtained from the swallow instants.
- Multi-layer perceptron algorithm can be applied as a supervised learning technique in all classification modules. Also, remember that swallowing is a sequential process and recurrent neural networks (RNN) or long-short-term memory (LSTM) are considered to be effective since they are preferred in cases where temporal behavior of data is significant. However, since they are deep learning methods, an increase in the data set may be required.
- As stated in most of the previous research, correct assessment of oral feeding skills is highly correlated with the synchronization of sucking, swallowing and respiration. Instrumental evaluation of all in a system will considerably improve the decision support mechanism.

REFERENCES

[1] C.-T. Chen, Y.-L. Wang, C.-A. Wang, M.-J. Ko, W.-C. Fang, B.-S. Lin, *et al.*, "Wireless monitoring system for oral-feeding evaluation of preterm infants.," *IEEE Trans. Biomed. Circuits and Systems*, vol. 9, no. 5, pp. 678–685, 2015.

[2] W. H. Organization, "Preterm birth." `http://www.who.int/news-room/fact-sheets/detail/preterm-birth`, 2018 (Accessed October 30, 2018).

[3] C. Lau, E. Smith, and R. Schanler, "Coordination of suck-swallow and swallow respiration in preterm infants," *Acta Paediatrica*, vol. 92, no. 6, pp. 721–727, 2003.

[4] H.-Y. Chang, P.-C. Torng, T.-G. Wang, and Y.-C. Chang, "Acoustic voice analysis does not identify presence of penetration/aspiration as confirmed by videofluoroscopic swallowing study," *Archives of physical medicine and rehabilitation*, vol. 93, no. 11, pp. 1991–1994, 2012.

[5] D. A. Ince, A. Ecevit, B. O. Acar, A. Saracoglu, A. Kurt, M. A. Tekindal, and A. Tarcan, "Noninvasive evaluation of swallowing sound is an effective way of diagnosing feeding maturation in newborn infants," *Acta Paediatrica*, vol. 103, no. 8, pp. e340–e348, 2014.

[6] R. K. Goyal and H. Mashimo, "Physiology of oral, pharyngeal, and esophageal motility," *GI Motility online*, 2006.

[7] Y. Fujiso, N. Perrin, J. Van Der Giessen, N. E. Vrana, F. Neveu, and V. Woisard, "Swall-e: A robotic in-vitro simulation of human swallowing," *PloS one*, vol. 13, no. 12, p. e0208193, 2018.

[8] C. Lau and R. J. Schanler, "Oral motor function in the neonate," *Clinics in Perinatology*, vol. 23, no. 2, pp. 161–178, 1996.

[9] F. Bu'Lock, M. Woolridge, and J. Baum, "Development of co-ordination of sucking, swallowing and breathing: Ultrasound study of term and preterm infants," *Developmental Medicine & Child Neurology*, vol. 32, no. 8, pp. 669–678, 1990.

[10] I. H. Gewolb, F. L. Vice, E. L. Schweitzer-Kenney, V. L. Taciak, and J. F. Bosma, "Developmental patterns of rhythmic suck and swallow in preterm infants," *Developmental medicine and child neurology*, vol. 43, no. 1, pp. 22–27, 2001.

[11] N. Amaizu, R. Shulman, R. Schanler, and C. Lau, "Maturation of oral feeding skills in preterm infants," *Acta Paediatrica*, vol. 97, no. 1, pp. 61–67, 2008.

[12] C. Lau and E. Smith, "A novel approach to assess oral feeding skills of preterm infants," *Neonatology*, vol. 100, no. 1, pp. 64–70, 2011.

[13] C. Lau, "Development of suck and swallow mechanisms in infants," *Annals of Nutrition and Metabolism*, vol. 66, no. Suppl. 5, pp. 7–14, 2015.

[14] G. D. Gramigna, "How to perform video-fluoroscopic swallowing studies," *GI Motility online*, 2006.

[15] B. Martin-Harris and B. Jones, "The videofluorographic swallowing study," *Physical medicine and rehabilitation clinics of North America*, vol. 19, no. 4, pp. 769–785, 2008.

[16] S. T. Almeida, E. L. Ferlin, M. A. M. Parente, and H. A. Goldani, "Assessment of swallowing sounds by digital cervical auscultation in children," *Annals of Otology, Rhinology & Laryngology*, vol. 117, no. 4, pp. 253–258, 2008.

[17] L. A. Newman, C. Keckley, M. C. Petersen, and A. Hamner, "Swallowing function and medical diagnoses in infants suspected of dysphagia," *Pediatrics*, vol. 108, no. 6, pp. e106–e106, 2001.

[18] T. U. Kim, W. B. Park, S. H. Byun, M. J. Lee, and S. J. Lee, "Videofluoroscopic findings in infants with aspiration symptom," *Journal of Korean Academy of Rehabilitation Medicine*, vol. 33, no. 3, pp. 348–352, 2009.

[19] K. E. Uhm, S.-H. Yi, H. J. Chang, H. J. Cheon, and J.-Y. Kwon, "Videofluoroscopic swallowing study findings in full-term and preterm infants with dysphagia," *Annals of rehabilitation medicine*, vol. 37, no. 2, pp. 175–182, 2013.

[20] S. E. Langmore, S. M. Kenneth, and N. Olsen, "Fiberoptic endoscopic examination of swallowing safety: a new procedure," *Dysphagia*, vol. 2, no. 4, pp. 216–219, 1988.

[21] S. E. Langmore, "History of fiberoptic endoscopic evaluation of swallowing for evaluation and management of pharyngeal dysphagia: changes over the years," *Dysphagia*, vol. 32, no. 1, pp. 27–38, 2017.

[22] S. Willette, L. H. Molinaro, D. M. Thompson, and J. W. Schroeder Jr, "Fiberoptic examination of swallowing in the breastfeeding infant," *The Laryngoscope*, vol. 126, no. 7, pp. 1681–1686, 2016.

[23] J. Reynolds, S. Carroll, C. Sturdivant, L. Ikuta, and K. Zukowsky, "Fiberoptic endoscopic evaluation of swallowing," *Advances in Neonatal Care*, vol. 16, no. 1, pp. 37–43, 2016.

[24] F. Weber, M. Woolridge, and J. Baum, "An ultrasonographic study of the organisation of sucking and swallowing by newborn infants," *Developmental Medicine & Child Neurology*, vol. 28, no. 1, pp. 19–24, 1986.

[25] D. T. Geddes, L. M. Chadwick, J. C. Kent, C. P. Garbin, and P. E. Hartmann, "Ultrasound imaging of infant swallowing during breast-feeding," *Dysphagia*, vol. 25, no. 3, pp. 183–191, 2010.

[26] K. Takahashi, M. E. Groher, and K.-i. Michi, "Methodology for detecting swallowing sounds," *Dysphagia*, vol. 9, no. 1, pp. 54–62, 1994.

[27] J. A. Cichero and B. E. Murdoch, "Detection of swallowing sounds: methodology revisited," *Dysphagia*, vol. 17, no. 1, pp. 40–49, 2002.

[28] E. W. Reynolds, F. L. Vice, J. F. Bosma, and I. H. Gewolb, "Cervical accelerometry in preterm infants," *Developmental medicine and child neurology*, vol. 44, no. 9, pp. 587–592, 2002.

[29] O. Amft and G. Troster, "Methods for detection and classification of normal swallowing from muscle activation and sound," in *Pervasive Health Conference and Workshops, 2006*, pp. 1–10, IEEE, 2006.

[30] E. W. Reynolds, F. L. Vice, and I. H. Gewolb, "Variability of swallow-associated sounds in adults and infants," *Dysphagia*, vol. 24, no. 1, pp. 13–19, 2009.

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

[32] M. Aboofazeli and Z. Moussavi, "Automated extraction of swallowing sounds using a wavelet-based filter," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 5607–5610, IEEE, 2006.

[33] M. Aboofazeli and Z. Moussavi, "Swallowing sound detection using hidden markov modeling of recurrence plot features," *Chaos, Solitons & Fractals*, vol. 39, no. 2, pp. 778–783, 2009.

[34] S. S. Shirazi, A. H. Birjandi, and Z. Moussavi, "Noninvasive and automatic diagnosis of patients at high risk of swallowing aspiration," *Medical & biological engineering & computing*, pp. 1–7, 2014.

[35] S. R. Youmans and J. A. Stierwalt, "An acoustic profile of normal swallowing," *Dysphagia*, vol. 20, no. 3, pp. 195–209, 2005.

[36] C.-T. Chen, L.-Y. Wang, Y.-L. Wang, and B.-S. Lin, "Quantitative real-time assessment for feeding skill of preterm infants," *Journal of medical systems*, vol. 41, no. 6, p. 95, 2017.

[37] F. L. Vice, O. Bamford, J. M. Heinz, and J. F. Bosma, "Correlation of cervical auscultation with physiological recording during suckle-feeding. in newborn infants," *Developmental Medicine & Child Neurology*, vol. 37, no. 2, pp. 167–179, 1995.

[38] E. S. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. L. Melanson, and M. R. Neuman, "Automatic detection of swallowing events by acoustical

means for applications of monitoring of ingestive behavior," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 626–633, 2010.

[39] O. Makeyev, P. Lopez-Meyer, S. Schuckers, W. Besio, and E. Sazonov, "Automatic food intake detection based on swallowing sounds," *Biomedical signal processing and control*, vol. 7, no. 6, pp. 649–656, 2012.

[40] Boersma, Paul & Weenink, David (2019). Praat: doing phonetics by computer [Computer program]. Version 6.0.49, retrieved 2 March 2019 from http://www.praat.org/.

[41] S. S. Meduri and R. Ananth, "A survey and evaluation of voice activity detection algorithms," 2012.

[42] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201–212, 1976.

[43] D. O'Shaughnessy, "Linear predictive coding," *IEEE potentials*, vol. 7, no. 1, pp. 29–32, 1988.

[44] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.

[45] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition.," in *Odyssey*, p. 7, 2010.

[46] J. P. Bello, "Low-level features and timbre." `https://www.nyu.edu/classes/bello/MIR_files/timbre.pdf`, 2016.

[47] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on selected areas in communications*, vol. 6, no. 2, pp. 314–323, 1988.

[48] S. S. Stevens and J. Volkmann, "The relation of pitch to frequency: A revised scale," *The American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.

[49] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, "Mfcc and its dynamic features," in *Spoken Language Processing*, 2001.

[50] R. O. Duda, P. E. Hart, and D. G. Stork, "K-means clustering," in *Pattern classification*, John Wiley & Sons, 2012.

[51] C. M. Bishop, "Mixtures of gaussians," in *Pattern Recognition and Machine Learning*, Springer, 2006.

[52] C. M. Bishop, "Mixture models and EM," in *Pattern Recognition and Machine Learning*, Springer, 2006.

[53] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[54] S. V. Vaseghi, "Baum–welch model re-estimation," in *Advanced Digital Signal Processing and Noise Reduction*, Wiley, 2008.

[55] D. G. Richard O.Duda, Peter E. Hart, "Support vector machines," in *Pattern Classification*, John Wiley & Sons, 2012.

[56] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[57] C. M. Bishop, "Sparse kernel machines," in *Pattern Recognition and Machine Learning*, Springer, 2006.

[58] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[59] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[60] J. Y. Gil and R. Kimmel, "Efficient dilation, erosion, opening, and closing algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1606–1617, 2002.

[61] S. W. Smith *et al.*, "The scientist and engineer's guide to digital signal processing," 1997.

[62] I. Pitas and A. N. Venetsanopoulos, “Median filters,” in *Nonlinear digital filters*, pp. 63–116, Springer, 1990.

[63] MATLAB, *version 7.10.0 (R2017b)*. Natick, Massachusetts: The MathWorks Inc., 2017b.

[64] W. Chai, *Automated analysis of musical structure*. PhD thesis, Massachusetts Institute of Technology, 2005.

[65] E. Peiszer, T. Lidy, and A. Rauber, “Automatic audio segmentation: Segment boundary and structure detection in popular music,” 2008.