

Use of Transfer Learning for Automatic Dietary Monitoring through Throat Microphone Recordings

by

Mehmet Ali Tuğtekin Turan

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Electrical and Electronics Engineering



KOÇ ÜNİVERSİTESİ

March 21, 2019

**Use of Transfer Learning for
Automatic Dietary Monitoring through
Throat Microphone Recordings**

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Mehmet Ali Tuğtekin Turan

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Prof. Engin Erzin (Advisor)

Prof. Çağatay Başdoğan

Prof. Levent M. Arslan

Prof. A. Murat Tekalp

Prof. Burak Acar

Date: _____

The necessary thing for anyone to be happy and contented as long as he lives is working for the ones who will come after him rather than working for himself. One can reach the true delight in life only by working for the existence and happiness of future generations.

Mustafa Kemal Atatürk

ABSTRACT

Use of Transfer Learning for Automatic Dietary Monitoring through Throat Microphone Recordings

Mehmet Ali Tuğtekin Turan

Doctor of Philosophy in Electrical and Electronics Engineering

March 21, 2019

Wearable devices and technologies in healthcare have been accelerating the development and integration of engineering approaches. Dietary monitoring is one challenging application among other healthcare services. Personal records are typically preferred ways of dietary monitoring. However, manual logging is highly biased and unreliable as individuals tend to underestimate their food intake. Automatic dietary monitoring (ADM) is an intelligent wearable coaching solution for this problem. The current research on ADM is shaped by the sensing devices and is fitted by machine learning approaches. In this thesis, we first define an ADM system using a throat microphone (TM) food intake sound recordings, where chewing and swallowing detection is presented. This system uses the TM sensor as a non-invasive transducer mounted on the neck, which is capable of delivering robust signal recordings for intelligent and unobtrusive food intake monitoring. Then, we investigate the use of transfer learning paradigm in depth to design an improved food intake detection and classification system. Although it is possible to reach abundant labeled food intake data recorded with traditional close-talk microphones (CM), data from TM modality is scarce. This creates a bottleneck in training deep architectures effectively using TM data. Recently, teacher/student (T/S) learning paradigm is introduced as a model compression framework, where it describes a class of learning methods for training a smaller student network by mimicking a larger teacher network. We propose a new domain adaptation framework in a heterogeneous setup based on T/S learning paradigm. The teacher network is trained over abundant high-quality CM recordings, whereas the student network takes TM recordings as input and distills deep feature extraction capacity of the teacher over a parallel CM and TM dataset. This allows the use of a significantly larger set of adaptation data, adds robustness to the resulting model, and significantly improves the performance of food intake detection.

ÖZETÇE

Gırtlak Mikrofonu Kayıtları Üzerinden Öğrenim Aktarımının Otomatik Diyet Takibi İçin Kullanımı

Mehmet Ali Tuğtekin Turan

Elektrik ve Elektronik Mühendisliği, Doktora

21 Mart 2019

Giyilebilir cihazların ve teknolojilerin sağlık hizmetlerinde kullanılması, mühendislik yaklaşımlarının geliştirilmesini ve entegrasyonunu hızlandırmaktadır. Diyet takibi, diğer sağlıkla ilgili servislere kıyasla zorlu bir uygulamadır. Günümüzde tercih edilen takip yöntemleri ise kişisel kayıtlara dayalıdır. Ancak bireyler genellikle gıda tüketimlerini hafife alma eğilimindedirler. Bu yüzden tutulan günlükler ve raporlar oldukça yanlış olup, gerçeği yeterince yansıtmamaktadır. Otomatik diyet takibi (ODT) bu soruna çözüm üretmek için akıllı ve giyilebilir bir yaklaşım önerir. ODT üzerine yapılan güncel araştırmalar, algılayıcı cihazlar üzerinden şekillendirilip, genelleksel makine öğrenimi kullanılarak eğitilirler. Bu konuda farklı bir yaklaşım önerilip, ilk önce çiğneme ve yutma algılamalarının sunulduğu bir ODT sistemi, gırtlak mikrofonu (GM) kayıtları üzerinden tanımlanır. Bu yaklaşım, GM algılayıcısını boyun üzerine monte edilmiş invazif olmayan bir yapı olarak kullanarak, gıda tüketiminin takip edilmesi için akıllı bir yapı sunar. Takip eden bölümlerde, gelişmiş tespit ve sınıflandırma sistemleri tasarlamak için öğrenim aktarımı olarak bilinen yeni bir yaklaşım derinlemesine araştırılır. Geleneksel konuşma mikrofonları (KM) ile kaydedilmiş bol miktarda etiketli tüketim verilerine ulaşmak mümkün olsa da, GM kayıtlarından gelen veriler azdır. Bu sorun, derin mimarileri eğitmede önemli bir eksiklik yaratır. Yakın zamanda öğretmen-öğrenci (Ö/Ö) öğrenim paradigması, büyük bir öğretmen ağını taklit eden daha küçük bir öğrenci ağını eğitmek için model sıkıştırma yöntemi olarak tanıtılmıştır. Bu tezde Ö/Ö eğitime dayanan yeni bir öğrenim aktarımı önerilmektedir. Öğretmen ağı, bol miktarda yüksek kaliteli KM kayıtları üzerinden eğitilirken, öğrenci ağı ise GM kayıtlarını girdi olarak alır ve paralel bir KM ve GM verisetinde derin öznetelikleri çıkarır. Bu, önemli ölçüde daha büyük bir verisetinin kullanılmasına izin verir. Diğer yandan Ö/Ö eğitimi, ortaya çıkan modele sağlamlık kazandırır ve gıda tüketiminin tespit ve sınıflandırma başarımını önemli ölçüde geliştirir.

TABLE OF CONTENTS

List of Tables	ix
List of Figures	x
Chapter 1: Introduction	1
1.1 Motivation	2
1.2 Use Cases	11
1.3 List of Publications	13
Chapter 2: Background	15
2.1 What is Food Intake?	16
2.2 Early Applications of Intake Monitoring	17
2.3 Manual Methods in Dietary Assessment	20
2.4 Automated Techniques in Dietary Assessment	23
2.5 Need for Wearable Systems	28
2.6 On-Body Sensor Research	30
2.7 Health Related Applications of On-Body Sensors	32
Chapter 3: Food Intake Monitoring	36
3.1 Classification of Eating Sounds	37
3.1.1 Methodology	40
3.1.2 Experimental Evaluations	46
3.1.3 Results	48
3.2 Eating Activity Detection	50
3.2.1 Methodology	53
3.2.2 Experimental Evaluations	57

3.2.3	Results	60
3.3	Discussion and Conclusion	62
Chapter 4:	Transfer Learning Approach	64
4.1	Introduction	65
4.2	Notations and Definitions	66
4.2.1	Homogeneous Approach	68
4.2.2	Heterogeneous Approach	73
Chapter 5:	Applications of Transfer Learning	79
5.1	Phoneme Recognition with Denoising Auto-Encoders	80
5.1.1	Methodology	82
5.1.2	Feature Extraction and Augmentation	82
5.1.3	Convolutional Neural Network Model	85
5.1.4	Hidden Markov Model Decoding	86
5.2	Knowledge Distillation in Food Intake Classification	87
5.2.1	T/S Learning for Domain Adaptation	94
5.2.2	Vanilla Distillation	96
5.2.3	Unsupervised Heterogeneous Domain Adaptation	97
5.3	Long-Short Term Memory for Sequence Modeling	99
Chapter 6:	Experiments on Selected Applications	102
6.1	Phoneme Recognition	103
6.1.1	Speech Datasets	103
6.1.2	Parameters for Feature Extraction and mSDA Calculation . .	104
6.1.3	CNN Architecture	104
6.1.4	Evaluation Metrics	105
6.1.5	Results	105
6.2	Food Intake Classification	106
6.2.1	Datasets	106

6.2.2	Experimental Setup	108
6.2.3	Baseline Performances without Domain Adaptation	110
6.2.4	Performance Analysis with Domain Adaptation	112
6.3	Discussion and Conclusion	117
Chapter 7:	Conclusions	119
	Bibliography	122



LIST OF TABLES

2.1	Error/Bias in intake estimates from the FFQ	23
2.2	Categories of wearable sensors and some sample research.	31
2.3	The application areas of on-body devices.	33
3.1	Time and frequency domain descriptors in the feature set	45
3.2	A summary of the food intake dataset	47
3.3	Classification accuracies of the selected feature sets.	50
3.4	Performances of the proposed food intake detection	60
3.5	Confusion matrix for the LOSO experiment	61
3.6	Confusion matrix for the 2-fold experiment	61
6.1	PER results for the conventional GMM/HMM system	105
6.2	FER and PER results for the CNN/HMM system	106
6.3	Baseline performances of the CNN, ConvLSTM and Bi-ConvLSTM topologies (%)	110
6.4	Performances with the proposed T/S Learning based domain adaptation for the student domain $\mathcal{D}^S$ (%)	112
6.5	Confusion matrix of the Bi-ConvLSTM network with domain adaptation for the LOSO experiment	113
6.6	Performances of the proposed domain adaptation using the Bi-ConvLSTM network with the VTLP augmentation (%)	116

LIST OF FIGURES

1.1	Wearable systems for continuous monitoring. Commercially available examples: a) Fitness trackers, b) BioStampRC, c) Sensing food dynamics	5
2.1	Overview of food intake mechanism with corresponded technologies and studies	16
2.2	Top row images show the ceiling light camera and a typical taken photo. The bottom-left image is showing the smart table surface. The HAPIfork is shown in the bottom-right picture.	19
2.3	Overview of dietary monitoring approaches	21
2.4	Example of image-based dietary monitoring systems. Sensing with mobile devices are in (a), (b), and (c). Sensing with a wearable camera is in (d).	25
2.5	A time line of remarkable wearable devices.	29
3.1	Block diagram of the proposed food intake classification system . . .	40
3.2	First iteration of sifting process (left), first IMF after repetition of the algorithm (right)	42
3.3	A sample empirical mode decomposition of a food intake signal . . .	44
3.4	Waveform and spectrogram representations of different food intakes .	47
3.5	The mRMR ranking of the feature descriptors over the EMD components. Darker colors represent higher rankings.	49
3.6	Block diagram of the proposed food intake detection system	54
3.7	Sample spectrograms of food intake sounds	55
3.8	CNN architecture of the food intake detection system	57

3.9	Sample images from the data collection sessions	58
4.1	Different schemes between traditional learning (left) and transfer learning (right)	66
4.2	Different feature transformation schemes, the symmetric mapping of the source and target domains into a common latent feature space is on the left, the asymmetric transformation of the source domain to the target domain is on the right.	67
5.1	Block diagram of the proposed phoneme recognition system.	83
5.2	The architecture of T/S learning using parallel data for the proposed heterogeneous domain adaptation	97
5.3	Inner structure of the ConvLSTM [Xingjian et al., 2015]	100
5.4	Bidirectional operations in the ConvLSTM module	101
6.1	Organization of the input speech features to ConvLSTM models . . .	109
6.2	Mel-spectrograms of the CM and TM recordings	111
6.3	Average LOSO accuracy performance as a function of weight γ	114

Chapter 1

INTRODUCTION

Eating is a universal activity that is strongly related to everyday human life. It is also one of the most important health determinants. Therefore, it is no surprise that understanding eating habits and its consequences at the individual and social levels is incredibly valuable. The approaches in this dissertation are motivated by applications of eating detection along four inter-related dimensions: population health, nutritional epidemiology, dietary self-monitoring, and patient care.

Every year, significant resources are invested to improve medical treatments, develop pharmaceutical drugs, and mitigate the effects of various diseases. However, less emphasis is placed on preventative healthcare. By ensuring that individuals exercise regularly and eat a balanced diet, it can be reduced that their chances of various health risks such as cancer, diabetes, and heart disease. Although public health measures encourage individuals to lead healthy lifestyles, quantifying diet and exercise has generally remained an unaddressable challenge.

Increasingly active research on wearable technologies and emerging consumer electronics indicates a wearable world ahead. This section provides a brief motivation on wearable computing, sensors, and devices. Wearable devices are altering many aspects of society, including, but not limited to, health care, fitness, sports, and entertainment. This introductory chapter examines the changes that wearable technology has brought to these areas. It discusses the difficulties and potential challenges that need to be addressed to advance wearable technology in terms of materials, wearability, power supply, security, communication, and embedded computing of wearable sensors and systems.

1.1 Motivation

Every day individuals are faced with multiple choices about what and how much to eat. Some of these choices are known to be associated with under-nutrition, overfeeding, or chronic disease. How much these choices increase the dietary risk factors or lead to the potential beneficial factors in an individual or in populations has been revealed through dietary assessment, often in large-scale epidemiological studies. Nutritionists employing dietary assessment have repeatedly emphasized the importance of these choices [Thompson and Subar, 2017].

Eating is one of the most fundamental human activities. Satisfying the hunger urge is essential for survival and sharing a meal has been one of the most enduring social practices for thousands of years. Because of the important role eating plays in our lives, it has been extensively studied. Anthropologists have investigated the relationship of eating behavior to culture and society and have claimed that learning how food is eaten is to learn how a society functions [Barthes, 2012]. Food consumption has been shown to be tied to rituals, symbols, belief systems and identities [Vermeir and Verbeke, 2006].

For several decades health researchers have also been deeply interested in studying eating habits and its impact on human health. It is now understood that good nutrition is vital for optimal growth and development, and prevention of disease [Go et al., 2005]. Dietary intake has been widely examined as it relates to cardiovascular disease, hypertension, obesity, diabetes, cancer, osteoporosis and many other medical conditions [Kearney, 2010].

Nutritional epidemiologists have typically relied on validated dietary assessment instruments driven by self-reported data including food frequency questionnaires and meal recalls [Wilde and Garvin, 2007]. Unfortunately, these instruments suffer from several limitations, ranging from biases to memory recollection issues [Jacobs, 2012].

In the beginning of the 1970s, the emergence of bio-markers has been a major development for efforts to understand and increase the reliability of dietary

assessment. Examples include 24h urinary nitrogen to validate protein intake [Huse et al., 1974], the PABA technique [Bingham and Cummings, 1983] for recovery of intervention-induced changes of specific food items, and doubly-labeled water [Westerterp et al., 1986] for validating the energy intake under conditions of energy balance. With regard to the latter, major insights came through doubly-labeled water energy expenditure studies that have revealed that under-reporting was endemic in dietary surveys [Livingstone et al., 1990], not method specific [Livingstone and Black, 2003] and mainly related to dietary fat intake in Western societies [Goris and Westerterp, 2000]. Conversely, over-reporting was identified, when caretakers reported on the patient's food intake [Simmons and Reuben, 2000].

Food intake aims at compensating energy expenditure, hence to attain a balanced metabolism. In the 20th century, the original challenge to acquire food became a commodity and many foods were designed with high energy content, oils, fat and caloric sweeteners [Tanumihardjo et al., 2007]. Concurrently, energy expenditure has decayed. This is primarily due to reduced physical activity, required to accomplish everyday tasks and work. It is reported a global rise in body fat, determined by the body mass index 1 (BMI) as consequence of energy imbalance [Malik et al., 2013]. BMI is used to identify overweight ($BMI > 25$), or more severe, obesity ($BMI > 30$). Both, overweight and obesity are a predispose for cardiovascular diseases, diabetes type 2 and further health risks, all eventually increasing morbidity [Visscher and Seidell, 2001]. For 2015, it is estimated a pandemic of 1.6 billion overweight and 400 million obese adults worldwide [Franz et al., 2015]. An even increasing trend was projected for 2015 that emphasizes a surge in child and adolescent obesity.

Chronic conditions account for more than 75% of health care expenses in the United States, and are the leading cause of deaths and disabilities [Dietz et al., 2016]. Preventive measures and better management can mitigate adverse effects associated with several chronic illnesses. Amongst the major conditions are obesity, eating disorders and diabetes, all of which are significantly affected by dietary behavior. Therefore, food intake monitoring is a promising research direction in the fight

against obesity, eating disorders, and associated health conditions.

Obesity alone is known to increase an individual's risk for type II diabetes, heart disease, high blood pressure, arthritis-related disability, stroke and some types of cancer [Smith and Smith, 2016]. Obesity affects more than 1 in 3 adults and more than 1 in 6 children and adolescents from ages 6 to 19 [Yang and Colditz, 2015]. For obesity prevention, monitoring physical activities (or energy expenditure) in daily living is only half the battle because food intake plays a notable role in maintaining energy balance. Weight gain (or weight loss) often results from an energy imbalance; over time, when people eat and drink more calories than they burn, the energy balance tips towards weight gain and vice versa [Vilar-Gomez et al., 2015].

On the other end of the spectrum from obesity is extreme weight loss caused by eating disorders such as anorexia and bulimia. Among modifiable health risk behaviors, regular physical activity and good nutrition from food intake can lower the risk of many of the top chronic conditions [Aune et al., 2016]. As seen from the examples in Figure 1.1, continuous monitoring using wearable systems has shown to be a reliable approach for activity and health tracking. Methods have been developed for accurate and objective characterization of physical activity [Lara et al., 2013], as well as estimating energy expenditure [Troiano, 2006]. Meanwhile research efforts are in progress for monitoring dietary behavior, and there is currently no accurate and non-obtrusive means to objectively monitor food intake in daily-living conditions [Thomaz et al., 2013].

Although many other food-related health problems exist, among them, obesity is probably the most common one which decreases the quality and duration of life and increases global healthcare costs. People's eating behavior has been shown to affect health issues. For example, a higher meal frequency has classically been associated with a lower incidence of obesity while eating outside the home, consuming fast food and eating takeaway food tend to be associated with obesity.

The prevalence of obesity in the US population (aged 20 years and over) increased from 14% in 1980 to 23% in 1994 and reached 30% in 2000 [Wang et al., 2006]. Moreover, by 2000 the ratio of overweight US-citizens exceeded 64%. In 2004, 17%



Figure 1.1: Wearable systems for continuous monitoring. Commercially available examples: a) Fitness trackers, b) BioStampRC, c) Sensing food dynamics

of US-children and adolescents aged 2 to 19 years were overweight [Skinner et al., 2016]. Eating disorders now affect up to 30 million people in the U.S. [Brownell and Walsh, 2017]. National prevalence among adults in Europe ranges from 8% in Switzerland and 10% in Italy and the Netherlands to 25% in England and nearly 30% in Greece and Croatia. In 2008 about 21% of the German adult population was obese [Kurth and Schaffrath, 2007]. Estimations for economic cost of obesity range between 2% and 8% of total healthcare costs in several developed countries. While these estimates are conservative, obesity represents one of the largest cost items in national healthcare budgets [Lehnert et al., 2013]. Moreover, overweight and obesity is not restricted to high-industrialized regions and is even faster growing in developing countries. In 2002 about 15% of the Chinese citizens were overweight [Gu et al., 2005]. Therefore, fighting these epidemic dimensions is a critical challenge for the success of our species.

Besides energy balance, food intake provides unique access to nutrients that cannot be sufficiently synthesized by the body, such as vitamins, minerals and water. This explains the enjoyable stimulus and motivation for a diversified diet composition. It warrants sufficient amount of all required nutrients because any form of eating disorder is detrimental for health [Heath, 2002]. Several further

eating disorders exist, including crash dieting, anorexia nervosa, binge eating, bulimia and orthorexia [Brownell and Walsh, 2017]. Hence, there are many aspects influencing individual food intake, including genetic and physiologic as well as psychological and social constraints [Jéquier and Tappy, 1999]. The result is an individual eating behavior described by specific food choice and restraint, portion size, energy intake and meal intake frequency. Eating behavior is reported in temporal resolutions ranging from individual snacks and meals every day to averages over several years, depending on the type of investigation [Jakubowicz et al., 2013].

These facts have fostered the emergence of many approaches to monitoring diet, which record and analyze the content of each meal, as well as the general eating habits of a person. However, most currently used self-reporting techniques demonstrate widespread underestimation of food intake, related to diverse psychological implications [Black and Livingstone, 2003]. Under-reporting of dietary energy intake is common even in non-obese adults [Poppitt et al., 1998].

To date, most investigations and weight loss programs assess daily eating behavior (intake schedule, food composition, amount and energy content) with the help of questionnaires [Fernández-Ballart et al., 2010]. Questionnaire assessments in the form of self-reports could capture eating behavior in the required temporal resolution and information detail. However, they fail due to the burden of manual logging [Rabbi et al., 2015]. All diet assessments are either laboratory-based or require a considerable effort by the respondent.

Other factors like a subject's awareness of food intake (particularly with snacks), literacy level, memory and perception capabilities can also affect self-reporting [Jakubowicz et al., 2013]. Thus, self-reported food intake data should be interpreted with caution [Konttinen et al., 2010]. Because of this, many researchers have started developing affordable and non-invasive solutions to help users measure and analyze their food intake more precisely. This idea has also inspired the Food Scanner prize, a one-million Euro competition organized by the European Commission [Rateni et al., 2017].

Dietary reporting has not been impossible to attain. Under the guidance of dietitians, subjects partly learned to report their own food intake with full accuracy [Goris and Westerterp, 1999]. On the negative side, even in the hands of the most experienced nutritional epidemiologists, inaccuracies in dietary assessment may remain [Freedman et al., 2014]. Furthermore, even if reports were accurate, some research questions have been too specific when using traditional dietary epidemiological methods. For instance, when the range of intake distributions is small, large-scale studies probably cannot find diet–health associations.

Some other studies admit that accuracy of reported or recalled absolute intakes fails, yet dietary patterns can be recognized via traditional dietary assessments [Yon et al., 2007]. However, even if dietary assessment is limited by some inherent weaknesses, it should not be discredited and rejected for research purposes. Indeed, the information should be used with caution, should not be over-interpreted, and would be advantageously used with other supporting techniques [Satia-Abouta et al., 2002]. A multi-step process should be used to reinforce the outcome of a dietary assessment. Therefore, objective methods are projected to add useful information that would compensate for shortcomings of traditional dietary assessment.

Although early methods for dietary assessment have evolved and improved, they still suffer from a lack of accuracy in reporting due to memory failure, response bias, and the daily variability in diet. The meal-to-meal or day-to-day variations in food intake make it difficult to capture usual methods. Dietary reporting is inconvenient, and individual biases such as body image, weight consciousness, dietary restraint, and social expectations, may influence it [Tylka et al., 2014].

In the face of these aspects, research have been diverged on dietary assessment and its traditional tools. Specifically with respect to the food intake, recent studies has moved away from the use of self-reports as a research tool. They base this on the evidence that self-reports of energy intake are so poor that they are wholly unacceptable for scientific research on developing objective measures of energy

intake [Dhurandhar et al., 2015]. Reduction of the subject's burden, increment in the acceptability and compliance, applicability to diverse population groups will deliver higher data quality, consistency, and reliability of the assessments. These are expected to lead greater success in the assessment of dietary intake [Forster et al., 2016].

With increasing health awareness, and numerous health-related advisers, the question remains as to how detailed dietary assessment still can become more relevant. Current research questions are asked in the perspective of the leading causes of death: risks for cardiovascular diseases, chronic metabolic disorders, and cancer [Kina-Tanada et al., 2017]. Personalized nutrition is also a current focus with dietary assessment methods in an attractive format and with appropriate feedback that are tailored to the individuals [Srinivasan et al., 2017]. To these ends, investigators are beginning to combine dietary assessment with a user's profile, so that subjects may reduce their risk of dietary diseases as part of personalized wellness.

These shortcomings can be partly or completely circumvented during the near future. Improved sensors, and portable electronic technology are all under investigation. Photography may help to add information about ingredients, food preparation, and food consumption environment. Three-dimensional scanning of foods may be obtained in restaurants and shops to improve estimates of serving sizes, and a chip to measure nutrients in the blood is on the technical wish list [Sun and Saenko, 2016]. At the same time, updates of food composition databases and harmonization across cultures for international studies will receive attention [Dwyer et al., 2018]. However, these will not solve automation problem of the dietary assessment.

Wearable technology has experienced explosive growth in recent years. According to PricewaterhouseCoopers, 21% of surveyed American adults owned at least one wearable device in October 2014, but only one and a half years later, in April 2016, this figure had grown to 49% [Chang et al., 2016]. Smart watches, smart glasses, fitness bands, and other wearable devices have become common

accessories in daily life. They provide smart sensing, multi-modal interfacing, and monitoring capabilities in addition to wearability, data input, data recording, and communication and has many potential applications [Sung et al., 2005]. Therefore, wearables provide a potential perspective in dietary applications, and future challenges.

Recently, several methods have been proposed to improve diet monitoring by automating the whole task or part of it in order to improve its accuracy and make it more user-friendly and practical. This research field is relatively young; because of this, research efforts mostly represent sporadic attempts targeting different aspects of diet monitoring [Sundaravadivel et al., 2018]. Usually, each study describes its own unique approach and does not follow any well-established trend or research line. Nevertheless, patterns of similarity can be detected from a wider viewpoint and some relevant methods can be identified.

Automated monitoring of eating behavior is the prerequisite for research on disease intervention and epidemiology as well as in deployed prevention, such as weight loss coaching programs [Stein and Brooks, 2017]. In order to systematically reduce disease risks, these programs target a modification of accustomed lifestyle. However, this is a tough challenge for the individual. It requires continuous, potentially life-long, everyday support and coaching. To support the coach and individual with actual information, eating behavior reporting must provide a similar temporal resolution, thus requires tracking of every individual meal intake. Such actual information is particularly vital to adapt feedback in coaching programs and has been identified to improve success rates [Aschbrenner et al., 2016].

Current studies on eating behavior and weight loss consider typical intervention periods of six months. However, only 20% of the individuals that initially lost at least 10% of their weight, can maintain the new weight one year after discharge [Soleymani et al., 2016]. Novel tools and technical solutions are sought that alleviate the individual from manual food intake logging. The vision for such solutions is to provide eating behavior information in the conceptual

quality of daily self-reports. This is the goal of automatic dietary monitoring (ADM). These solutions will remove the inter-individual estimation error and increase user compliance in interventions. Moreover, they would permit novel risk-prevention programs through long-term personalized coaching that is clearly infeasible using manual monitoring.

Many factors affect and influence dietary choices and overall health of a person. Some of these are individual factors (age, gender, and physical activity patterns), environmental settings (school, workplace and recreational facilities), sectors of influence (government and health care system), as well as social and cultural norms that govern thoughts, beliefs and behavior. Over the last 10 years, a large body of research has aimed at automating the task of food intake monitoring [Bettadapura et al., 2015]. This research has been possible thanks to advances in mobile, wearable and sensing technologies. Despite significant progress, most proposed systems have required individuals to wear deficient and scarce devices. These form-factor requirements have severely limited the immediate practicality of automated food intake monitoring in health research.

Another factor that has hampered progress in automated nutrition monitoring has been the way the recognition problem has been represented. There are at least two key technical challenges in building a fully automated food intake monitoring system: (1) recognizing when an individual is performing an eating activity, and then (2) inferring what and how much the individual eats. Historically, methods devised to automate dietary tracking have largely ignored the distinction between these challenges. This is an important reason why previous automated dietary assessment research efforts did not meet expectations in terms of practical applicability and deployment.

This thesis tries to provide a wide perspective of the research field, structuring the discussion according to the main sub-problems that have to be tackled by the methods, describing the main approaches used to solve them and the main challenges which remain open. Following sections provide an analytical discussion that highlights the main weaknesses or strengths of the approaches, both from a

general viewpoint and, when relevant, individually for the methods described therein. My work is particularly concerned with the challenge of eating detection. The aim of this dissertation is to answer the question that it is possible to automatically detect and classify consumed foods via on-body wearable sensors.

1.2 Use Cases

Nutritional epidemiology forms the backbone of public health, guidelines and even agricultural subsidies. One of the key reasons why researchers are interested in how people eat is to elucidate the mapping between dietary habits and disease. Finding out what people eat and the broader context of eating activities has been of interest to epidemiologists for many decades [Hatori et al., 2012].

One disease, cancer, has been extensively studied alongside dietary impact. Cancer is considered a chronic disease of the genome that may be influenced at many stages by nutritional and metabolic factors. It is estimated that up to 80% of colon, breast, and prostate cancer cases and one-third of all cancer cases may be influenced by diet and associated lifestyle factors [Go et al., 2005]. The source of divergent findings in nutritional epidemiology stems in large part from the use of flawed measurement tools.

Like health studies, epidemiological research is also based on self-reported data. In fact, population health data sometimes drive explorations in the space of nutrition. And as previously mentioned, self-report based instruments have many flaws, which are detailed in Chapter 3. Recently, there has been a strong sentiment in the health research community that more resources need to be allocated towards the development of more objective and precise measures, which includes the ability to detect eating activities [Dhurandhar et al., 2015].

The need for improved dietary tracking is also shared by individuals interested in meeting health goals. Recently, health concerns linked to dietary behaviors such as obesity and diabetes have fueled demand for dietary self-monitoring, one of the most effective methods for weight control [Burke et al., 2011]. It is characterized by systematic self-observation, periodic measurement, and recording of target

behaviors with the goal of increasing self-awareness. However, adherence to dietary self-monitoring is poor and generally wanes over time, even with modern smartphone-based systems [Cordeiro et al., 2015]. Individuals must remember to log meals and snacks throughout the day, and then manually record eating activities, a tedious and time-consuming task.

Semi-automated food journaling, a technique that hinges on eating detection, is a promising new approach where the food tracking task is split between individuals and an automated system, thus reducing the burden of self-monitoring while keeping individuals involved in the process. In essence, it is an attempt to reach a compromise between manual and automatic nutrition tracking. The key aspect of the approach is the splitting of the food journaling task into two sub-tasks. The first sub-task, which is fully automated, centers on detecting when an eating activity is taking place. Examples of eating activities include breakfast, lunch, dinner, and snacks. Upon eating detection, the other sub-task is triggered, requiring individuals to manually provide some information pertaining to what was consumed.

Once eating is taking place and the eating activity has been detected, several courses of action could be pursued to prompt the individual for more information. In one scenario, the individual's smart-watch could softly vibrate to remind and nudge the individual to add an entry to a food log. In some cases, it might be undesirable or not socially acceptable to document a meal while it is taking place. Instead, the individual could receive a text message later in the day as a reminder to log at an opportune time in the near future.

While healthy eating habits are important in the prevention of a large number of medical illnesses for the general population, it is particularly critical for the elderly and individuals with chronic diseases. For older adults, poor nutritional intake is linked to increased morbidity and mortality due to energy deficiencies, low-body mass, cognitive decline, and many other factors [Marshall et al., 2001].

Poor dietary habits are also common for individuals with chronic diseases such as mental illnesses. For example, individuals with schizophrenia have a 20%

shorter life expectancy than the population at large and are vulnerable to lifestyle diseases including diabetes, coronary heart disease, and hypertension [Marder et al., 2004]. To make matters worse, some of the medications used to treat schizophrenia have been associated with weight gain, the onset of diabetes and other problems. The combined effect of these risks suggests that physical activity and nutrition monitoring as a means of health promotion would be beneficial to this population.

1.3 List of Publications

Paper [1*] M. A. Tuğtekin Turan and Engin Erzin. “Improving Phoneme Recognition of Throat Microphone Speech Recordings Using Transfer Learning.” *submitted to Elsevier’s Speech Communication*, 2019.

Paper [2*] M. A. Tuğtekin Turan and Engin Erzin. “Automatic Dietary Monitoring using Domain Adaptation for Throat Microphone Recordings.” *submitted to IEEE Transactions on Neural Networks and Learning Systems*, 2019.

Paper [3] M. A. Tuğtekin Turan and Engin Erzin. “Monitoring Infant’s Emotional Cry in Domestic Environments Using the Capsule Network Architecture.” *INTERSPEECH*, 2018.

<https://doi.org/10.21437/Interspeech.2018-2187>

Paper [4*] M. A. Tuğtekin Turan and Engin Erzin. “Detection of Food Intake Events From Throat Microphone Recordings Using Convolutional Neural Networks.” *IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

<https://doi.org/10.1109/ICMEW.2018.8551492>

Paper [5*] M. A. Tuğtekin Turan and Engin Erzin. “Empirical Mode Decomposition of Throat Microphone Recordings for Intake

Classification.” International Workshop on Personal Health and Health Care, ACM Multimedia, 2017.

<https://doi.org/10.1145/3132635.3132640>

Paper [6] M. A. Tuğtekin Turan and Engin Erzin. ”Source and Filter Estimation for Throat Microphone Speech Enhancement.” IEEE Transactions on Audio, Speech and Language Processing (TASLP), 2016.

<https://doi.org/10.1109/TASLP.2015.2499040>

Paper [7] Johannes Abel et al. ”A Subjective Listening Test of Six Different Artificial Bandwidth Extension Approaches in English, Chinese, German, and Korean.” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

<https://doi.org/10.1109/ICASSP.2016.7472812>

Paper [8] M. A. Tuğtekin Turan and Engin Erzin. ”Synchronous Overlap and Add of Spectra for Enhancement of Excitation in Artificial Bandwidth Extension of Speech.” INTERSPEECH, 2015.

https://www.isca-speech.org/archive/interspeech_2015/i15_2588.html

[*] Those related to this thesis are shown with an asterisk.

Chapter 2

BACKGROUND

This chapter summarizes the previous approach and the most important achievements in the ADM problem. Specifically, on-body sensing solutions for diet monitoring are analyzed, their performance is summarized, and the results of different algorithms are presented.

Automated, sensor-based methods are being explored in research to provide a more accurate and reliable alternative for dietary monitoring. From the human body perspective, it is important to first understand specific activities related to eating that can be monitored using a sensor-based approach. Some of these activities are swallowing, chewing, intake gesture, gastric activity, cardiac responses etc. On the other hand, wearable sensors are attached to the user in one or more locations and are capable of ubiquitous sensing. For this reason, wearable systems are sometimes the preferred means for continuous dietary monitoring.

On-body locations such as in-the-ear and on-the-neck are familiar with wearable locations for headphones/hearing aid systems and necklaces, respectively. In this chapter, we explore several factors that can affect the quality of a recorded signal from the monitoring perspective including sensor type(s) and design, on-body sensing location, recording channels, and recording environment. These systems have been developed using single- and multi-sensor approaches for single- and multi-location on-body utility. The chapter also presents a discussion of limitations and an outlook, indicating open challenges and new research directions.

2.1 What is Food Intake?

Food intake is essential in energy balance which is represented by calories. It control the energy homeostasis, the match of energy intake and energy expenditure, which have shown to affect obesity, highlighting its importance on human health ([Schwartz et al., 2000]). Therefore, energy expenditure is an important variable in maintaining optimal health.

The intake process is divided into two processes: mechanical digestion and chemical digestion (see Figure 2.1). One type of chemical digestion is protein consumption, in which the body breaks down dietary protein into usable amino acids, then makes this nutrient available to the cells in support of muscle maintenance, immune function, hormone synthesis, red blood cell formation, and tissue repair. One commercial product for protein digestion monitoring is the ITSI Biosciences kit [Arun et al., 2012], yet wearable sensors do not contribute much to this area. We here focus more on the mechanical digestion process, since it encloses more inputs, which can be detected via sensors (e.g. vision, sound or pressure).

Mechanism of the bite is recognized as the very first stage in food intake. It involves two main steps: movement of wrist and hand to transport food to the mouth and movement of jaw muscles to take a bite. Wrist movements are typified by characteristic rolling motions (pronation and supination) to deliver food from cutlery to mouth. Following this is the jaw movements for the mouth to obtain

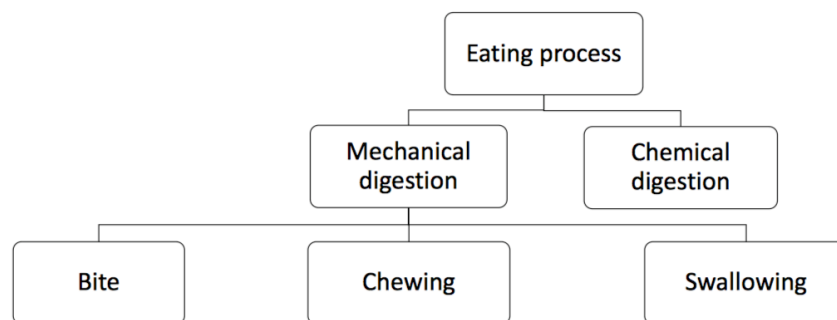


Figure 2.1: Overview of food intake mechanism with corresponded technologies and studies

the food. The lateral pterygoid muscle is responsible for the abduction of the jaw (opening mouth), while temporalis, masseter and medial pterygoid muscles assist the adduction (closing mouth, moving mouth side to side and grinding teeth) [Vu et al., 2017].

Mechanisms of chewing (mastication) involves the physical degradation of food. This digestion occurs mainly from the mouth, passing through the pharynx and esophagus. It starts in the oral digestion with the mastication process, to produce bolus, which is later swallowed down into the esophagus. In greater detail, during the oral process, the ingested food is moved from the front of the mouth to the teeth so that it can be broken down by crushing and grinding with the teeth. The jaw movements in this stage are similar to those in the mechanisms of bite with continuously indicative motions of lateral pterygoid and temporalis muscles to open and close the mouth [brown et al., 1998].

Mechanisms of swallowing begin with the voluntary phase. The bolus, after being formed by folding and manipulating the mixture of food particles and saliva with the tongue, is moved into the back of the tongue to prepare for swallowing. In the involuntary pharyngeal phase, the swallowing reflex is triggered to open and close the upper esophageal sphincter. During this phase, pharyngeal constrictor muscles push the bolus into the esophagus via a stripping action. The last phase, the involuntary esophageal phase, consists of continuous peristalsis by sequential contraction of the superior, middle and inferior pharyngeal constrictor muscles of the esophagus to push the bolus to the stomach [Buettner et al., 2002]. These movements of smooth and striated muscles in pharynx and esophagus play a vital role in the source of input for wearable sensors.

2.2 Early Applications of Intake Monitoring

Food intake becomes extremely crucial for patients in a hospital setting, and an on-going concern for medical centers. Only 25% of patients receive a sufficient amount of energy and protein, and 20% exhibit nutrition risks [Singer et al., 2014]. In addition, hospital nutrition management is lacking in several aspects including

planning and managing nutritional care, nutrition education among medical staff, limited influence on patient behavior. These are all factors beyond the control of wearable sensors, however, monitoring can help create an objective and reliable measure of food intake, facilitating a solution that helps solve hospital nutrition management.

The eButton from Sun et al. is a creative solution among existing approaches, utilizing a camera embedded in a button for capturing images of food on a plate [Sun et al., 2014]. The system also includes a smartphone to act as a control and information gateway. This device is capable of classifying food types, estimating the volume of the food and exporting the output of the calories and nutrients of the meal. The ability of this device is satisfactory for use in a hospital. Nutritionists and nurses can place it on patients' chests for automatic food data collection. This data can be stored in a local database for monitoring nutrition intake.

Eating behavior is affected by habit, whereby some participants exhibit night eating, evening overeating, as well as other external aspects, such as environmental factors, and social factors such as stress in daily life. Hence, an objective system is necessary for controlling food intake, as well as human weight.

An experimental approach is to use microwave Doppler motion sensors. Tanigawa et al. explored the Doppler effect in their system to sense the Doppler signal of mastication produced from vertical jaw movements [Tanigawa et al., 2008]. The distinct moving speed of the jaw during mastication is determined from the equation of relationship between Doppler frequency and moving speed.

Liu et al. merges both the acoustic (microphone) and the visual (camera) systems to detect eating sounds in applications for obesity [Chen et al., 2007]. With the assistance of the camera, this system is capable of detecting food portion by a time during the eating process through image processing of snapshots of the meal over time. Integrating this with information regarding food type, it can provide helpful data related to eating behavior, which directly impacts obesity.

Chang et al. embedded a weighing sensor and radio-frequency-identification (RFID) reader as a two-layer sensor surface beneath a dining table, to trace the



Figure 2.2: Top row images show the ceiling light camera and a typical taken photo. The bottom-left image is showing the smart table surface. The HAPIfork is shown in the bottom-right picture.

weight of food intake [Chang et al., 2006]. Based on this system, it is able to track food consumption by measuring the weight of the food content, along with the food movement patterns detected using the RFID antenna. More novel research using pressure sensors was conducted by Zhou et al. which is basically a pressure sensor with the upper layer serving as both protection and a dining surface. Specific eating actions are detected when someone dines on this tablecloth. The sensor is less of a burden than the one in Chang et al. However, it is capable of detecting only eating movements without either estimation or food classification.

Maekawa built a system for automatically photographing meal eating activities from a camera mounted on a dining room ceiling light (see the figure below) [Maekawa, 2013]. According to the author, one of the difficulties of relying on individuals for documenting meals is that people often forget to take pictures while eating. A camera was configured to turn on and off with the ceiling light under the assumption that eating always takes place with the lights on.

Kadomura et al. explored a sensor-embedded fork around an interactive mobile application with the goal of monitoring and possibly modifying a child's eating behavior [Kadomura et al., 2013]. The fork was instrumented with motion sensors

for detecting changes in eating behavior state and a single-pixel color sensor to determine food colors. By tracking different foods by color, the system attempted to encourage children to eat a variety of food items. Another instrumented utensil is the HAPIfork [Kojima et al., 2016], shown in Figure 2.2. It was designed to sense and control the pace of eating, delivering vibrations when it identifies that the person is eating too fast. Fluid intake tracking through specialized and instrumented cups has also been a focal point of researchers. Lester et al. developed a method that uses optical, ion selective electrical pH, and conductivity sensors to sense and classify liquid in a cup.

Early applications generally focused on food journaling or nutrition tracking for understanding eating behavior or research purposes in medical usages. However, more recent studies mainly investigated the monitoring idea to be used for dietary assessment. Regardless of the method of choice, the primary goal was to quantitatively track food intake parameters in an effort to encourage healthier dietary behavior.

2.3 Manual Methods in Dietary Assessment

From the aforementioned influencers, monitoring food intake is amongst behaviors with the strongest evidence shown to have a positive impact on weight management. Figure 2.3 shows a summary of dietary monitoring approaches in the literature and in practice, including manual and automated methods. Automated methods can be divided into fully-automated and semi-automated, both of which can be monitored using wearable, hand-held and environmental systems. Wearable monitoring systems can be single-sensor or multi-sensor devices designed for single- or multi-location on-body use.

Wearable nutrition monitoring devices are becoming a subject of interest in the research community. However, the most pervasive techniques today are still manual record-keeping approaches. Although these techniques are simple and affordable, they're associated with user burden and poor accuracy. Furthermore, although manual techniques are often conducted without electronic technology,

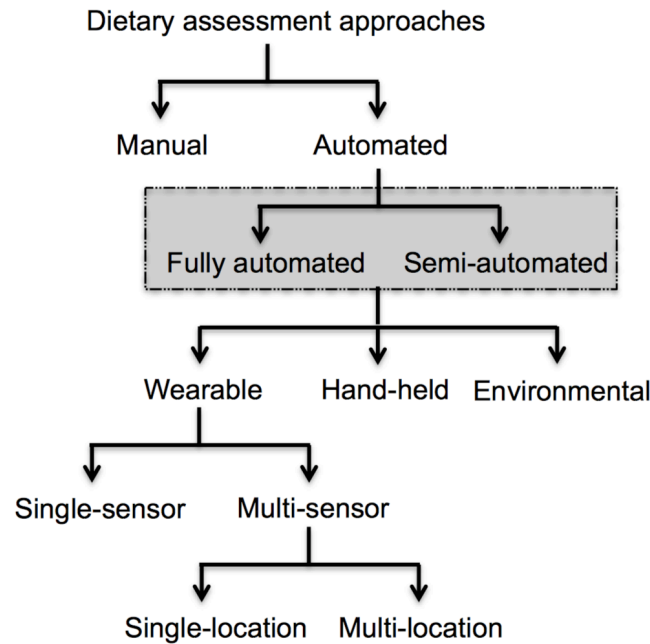


Figure 2.3: Overview of dietary monitoring approaches

several recently proposed applications, such as MyFitnessPal and SparkPeople, provide data logging capabilities and other resources for weight management.

A simple, pervasive method of monitoring diet is the multipass 24-hour dietary recall method, which is based on the data that patients provide at the end of a randomly selected day. Each individual gives an oral or written report including the amount and type of food eaten during the day, which is used to calculate total food intake. This approach measures food intake in a reasonably quantitative manner but with a significant error because people do not recall the exact amount of food they have eaten. Experimental data shows that food intake is usually reported with error, and measurement variance also depends on the patient's experience with the system.

Food records are generally not impacted by the accuracy of a subject's memory. They typically require individuals to note their eating habits during or immediately after a meal. Concerns include patient compliance and the difficulty that untrained individuals face when accurately assessing portion size. For example, caloric density varies based on the cooking method and other factors that

are not necessarily visually apparent [Rusin et al., 2013]. Moreover, manually recording each meal can be tedious, and many individuals will be unwilling to complete food records for extended periods of time.

Self-report methods require literate and motivated subjects. The added burden these methods place on subjects could be a reason for the decline in quality of food records relative to the number of days recorded. In addition, the process of actively recording food intake can cause subjects to change their eating patterns and this can lead to records containing an inaccurate representation of subjects' normal food intake. Independent of the self-reporting method used, underreporting by up to 50% is pervasive and this negatively affects efforts towards improving food intake habits and weight management [Gemming and Mhurchu, 2016].

A third method for manually assessing dietary intake is to use a food frequency questionnaire (FFQ), in which individuals specify their rate of consumption for various food items over the long term. Nutritional intake can subsequently be assessed by summing various food types provided within the list. This technique is inexpensive to administer and insensitive to recent changes in diet. Furthermore, it is less time-consuming than food records because it is not intended to be completed on a daily basis. However, FFQs are typically less accurate than other techniques. This is often a result of several factors, including incomplete lists of food, poor user compliance, and frequency and serving size errors.

Doubly labeled water (DLW) is another manual method often used to determine the validity of tools designed to measure energy intake. When using the DLW method, subjects are given a form of "labeled" water that includes elements such as deuterium and oxygen-18, which can be measured by sampling saliva, urine or blood to estimate metabolic rate [Murakami et al., 2016]. This method provides an accurate measure of a free-living subject's total energy expenditure which can be equivalent to energy intake in weight-stable individuals. Due to the high cost and sophisticated technology associated with DLW, its use to date is not suitable for personal purposes as the need for continuous food intake monitoring and DLW is confined to research laboratories.

Despite the use of these self-report methods for several decades, observations have shown that people tend to forget items that were eaten, underestimate large portion sizes, overestimate small ones and, in general, be susceptible to a large variety of errors and biases, some of which are shown in Table 2.1. Recently it has become possible to measure the accuracy of dietary recalls, records and FFQs thanks to the doubly-labeled water technique. Findings confirmed the weaknesses of these assessment methods.

Table 2.1: Error/Bias in intake estimates from the FFQ

Type of Error	Reason for Error
Memory	Unable to recall food consumption
Frequency judgment	Cognitive difficulty in providing information (low-literacy)
Question comprehension	Not able to understand which foods are being talked about
Response errors	Mistakenly codes incorrect frequency
Social desirability bias	Misrepresent dietary intake to please investigators

Manual approaches are simple and don't require user training for custom hardware operation. The initial cost is also relatively low because they don't require the purchase of any expensive devices. On the other hand, manual record keeping can be tedious, and long-term compliance rates are low. In addition, individuals often do not accurately remember what they have eaten or might deliberately misreport their eating habits.

2.4 Automated Techniques in Dietary Assessment

As previously stated, automated methods can be divided into fully-automated and semiautomated systems, which can further be divided into environmental, smart-object, handheld and wearable systems. Environmental sensing methods are fixed in predetermined locations and capture activities only inside these locations. Regions of interest for such systems are dining rooms, personal rooms in the home, and senior living centers.

Hand-held systems, primarily smartphones and mobile devices, have also been used towards dietary monitoring. In most cases, these systems allow for semi-automated monitoring because they rely on the user to trigger or initiate the recording process. An advantage of smartphone-enabled monitoring is that it builds on a system that users already carry around voluntarily and it can benefit from additional sensors embedded in the device such as camera, inertial sensors, and global positioning systems (GPS). On the other hand, these approaches can often identify specific foods, rather than reporting meal events, which can provide a more complete nutritional profile beyond volume or swallow count.

Figure 2.4 shows prototypes of image-based dietary monitoring systems. These systems rely on visual cues to supplement traditional self-report methods or to achieve automated monitoring using a single-sensor or multi-modal approach. In previous research, two primary platforms are used for image-based dietary sensing: 1) hand-held devices such as personal digital assistants (PDA), smartphones or tablets, 2) systems that include a camera [Wharton et al., 2014]. However, there is a potential drawback of inconsistent image quality from different smartphones/mobile devices for image-based methods. Sharp et al. highlight four dietary recording methods that use mobile phone platforms; namely, electronic food diary, food photograph-assisted self-administered 24-hour recall, food photograph analysis by a trained dietitian and automated food photograph analysis [Sharp and Allman-Farinelli, 2014].

The sound is a contextually-rich source of information that can be easily recorded using one of the simplest and most ubiquitous sensors; a microphone. Hence, a large body of work at the intersection of acoustic sensing and activity recognition has emerged over the last decade. Mobile phones have been explored extensively in auditory scene recognition and analysis. Rossi et al. implemented a system called AmbientSense as a mobile phone application for recognizing user context from ambient sounds [Rossi et al., 2013]. AmbientSense operated in two modes; it could perform audio recognition on the mobile device in real-time or by sending audio features for classification to a server. Given the difficulty of

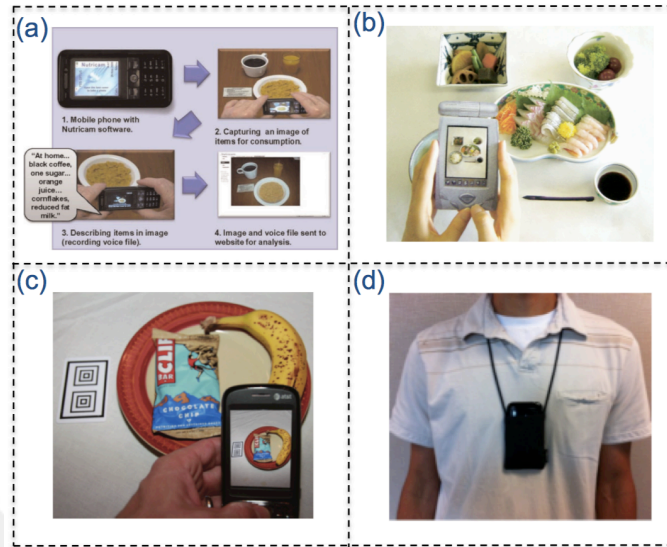


Figure 2.4: Example of image-based dietary monitoring systems. Sensing with mobile devices are in (a), (b), and (c). Sensing with a wearable camera is in (d).

collecting training data for a system like AmbientSense, Rossi et al. also examined the use of web-collected audio data to build a context recognition system. Through the compilation of audio data from the FreeSound database, they obtained practical recognition rates for the same classes studied in AmbientSense. Another implementation of an audio-centric activity recognition system on mobile devices is SoundSense [Lu et al., 2009]. It combines supervised and unsupervised machine learning techniques with a hierarchical classifier to perform varying levels of audio classification, and discover novel sound events specific to individual users.

There is no question mobile phones are ubiquitous, but most of the time they are inside pockets and purses. This raises an important question about practicability of mobile phone-based activity recognition systems. Franke et al. partially addressed this point by showing that a mobile phone can successfully infer ambient sounds even when inside clothing [Franke et al., 2009]. An alternative to mobile phones are wearable devices that are directly placed on the body; these are not limited by the constraint of being inside some other object (e.g. in a bag). A system that illustrates this approach is BodyScope whose goal was to explore how accurately a large number of activities could be recognized with a single acoustic microphone

[Yatani and Truong, 2012]. The system was able to recognize twelve activities in a lab study including eating, drinking, speaking, and laughing.

A food logging application based on the capture of audio and first-person point-of-view images was developed. The system processes all incoming sounds in real time through a head-mounted microphone and a classifier identifies when chewing is taking place, prompting a camera to capture a video of the eating activity. Gemming et al. categorize dietary image-based systems as either active or passive [Gemming and Mhurchu, 2016]. In the active sensing case, a user is required to initiate the meal-image recording process by, for example, taking a picture of the food per a specified protocol. Three common requirements were observed from previous literature as part of the protocol for meal-image capture in active sensing methods. First, meal-images should be captured at a specified angle, most commonly 45 degrees. Second, for quantifying of food intake, images of food selection before eating and leftovers after eating are necessary. Third, a visual reference object such as a PDA stylus, reference ruler or printed pattern is required to be included in meal-image pictures to facilitate estimation of parameters such as size, area, and color. The system in [Sharp and Sobal, 2012] was capable of projecting a light pattern on the meal-plate for use as a dimensional referent to calculate food portion size.

In the passive sensing case, a camera can be embedded in a wearable system or positioned to capture images (or videos) from a fixed location in an environment of interest such as a dining room. In these cases, the camera is automatically activated on a fixed time-basis or by detection of other activities like chewing [Liu et al., 2012]. Although passive sensing systems do not have the added burden of requiring users to capture meal-images, these systems have a higher probability of automatically capturing images or videos of other things/people in the scene and can, therefore, violate privacy. First-person images were captured every 30 sec. from a phone camera worn around the neck like a pendant. Due to the fixed timing for automatic image capture, such a system will require large memory capacity and high power consumption. In addition, due to possible privacy concerns of image-based passive sensing methods, [Thomaz et al., 2013] allowed an intermediary step for

users to review the entire image set and delete compromising or private images they did not want to share. Identification of eating episodes from ambient sound data was used to segment meal times in continuous video recording. Then, the video dataset was automatically scanned to identify and blur-out human faces captured during recording. Such privacy measures are particularly necessary for passive image-based methods.

The physiological approach is another type of automated systems that is comprised of two main sensor types: electroglottography (EGG) and electromyography (EMG) [Vu et al., 2017]. EGG sensors have long been used for the investigation of vocal cords' movements with the implementation of electrodes to detect electrical impedance across the neck near the larynx. This should not be confused with electrogastrography, which has the same abbreviation, but completely different specification (for gastric activity). EMG likewise employs electrodes but specifically targets skeletal muscles' activities. For such specifications, EMG and EGG are applicable to a variety of research in the field of automatic dietary assessment.

The EGG sensor utilizes the motion-induced variations of the electrical impedance, as the amplitude and/or phase change of high-frequency voltage, between two electrodes positioned on the larynx. Because of the characteristic of contacting the larynx, EGG has a high potential for wearable food intake monitoring. Farooq et al. proposed a testing scheme to evaluate the validity of practicing EGG in food intake detection by setting up the Laryngograph around the participant's neck during the experiment [Farooq et al., 2014]. The swallowing frequency signal will then be recorded by two electrodes embedded in the Laryngograph to detect eating events. The results afterward would be evaluated comparatively with the acoustic method using a microphone placed over the laryngopharynx.

The EMG has widely been employed in food intake monitoring research. Because of its efficiency in detecting chewing and swallowing signals, EMG focuses on mastication with the bolus size and hardness of the food intake, rather than

food classification. For example, Woda et al. employed EMG to scrutinize the effect of food hardness, bolus size, chewing cycles and sequence duration in specific food types and testing subject classes [Woda et al., 2006]. On the other hand, Kohyama et al. took into account mastication efforts for finely-cut foods with the use of the EMG [Kohyama et al., 2003]. A further research work of Kohyama et al. featuring EMG investigates more in-depth how chewing behavior is influenced by dental status and age with different types of food [Kohyama et al., 2007].

To summarize, automated techniques presented in this section are promising, especially in the analysis of intake patterns. However, it still needs more efforts and future work in order to expand their applications for food intake detection and classification problems.

2.5 Need for Wearable Systems

Since manual approaches are proven to be inaccurate and biased, we need a more objective and reliable system for dietary assessment. Wearable sensors are potential solutions since they do not rely on people's subjective judgments; moreover, wearable sensors provide real-time food intake monitoring. Data collected can be instantly synchronized with an information gateway, such as a smartphone application, for useful feedback. Another essential advantage for wearable sensors is creating an understanding of behavior in free-living people. It provides convenience for users, removing the burden of self-report, which can be helpful in medical applications, especially when the users are patients. The applications for wearable food intake monitoring systems are broad, which will be assessed in a later section of this paper.

Wearable technology has experienced explosive growth in recent years. According to PricewaterhouseCoopers, 21% of surveyed American adults owned at least one wearable device in October 2014, but only one and a half years later, in April 2016, this figure had grown to 49%. Smart watches, smart glasses, fitness bands, and other wearable devices have become common accessories in daily life. Wearable technology provides smart sensing, multimodal interfacing, and monitoring capabilities in addition to wearability, data input, data recording, and

communication and has a plethora of potential applications.

In addition, the accuracy of the automated wearable sensor approach is able to be improved by limiting the unwanted effects of external factors such as noises from the environment. For instance, the sensor-based approach requires denoising of the signal, which can even fully eliminate noises embedded in the collected data. With advancing technology, researchers have been able to develop satisfying denoising signal processing techniques. Ervin et al. succeeded in developing a procedure to denoise dual-axis swallowing accelerometry signals with significant improvements in comparison to previous approaches.

Historically, as shown in Figure 2.5, possibly the first successful wearable device in medicine might have been the hearing aid. The first commercial digital hearing aid was created by the Nicolet Corporation in 1987. It consisted of a pocket processor and a behind-the-ear transducer. Today, hearing aids have been further miniaturized and can be conveniently fitted in the outer-ear bowl or in the canal.

In 2004, using smart textiles and microelectronics, the CuteCircuit company introduced the KineticDress, which could change its patterns following the wearer's movements. Later on, CuteCircuit created several other smart garments, which could sense emotions, receive phone calls, or display images and texts. Also in 2004, the GoPro company introduced its first action camera. Wearable action cameras are used mostly for outdoor sports or law enforcement. Attached to helmets, armbands, or clothes, these devices can provide a first-person perspective and allow continuous capture of actions, while keeping the user's hands and vision free.

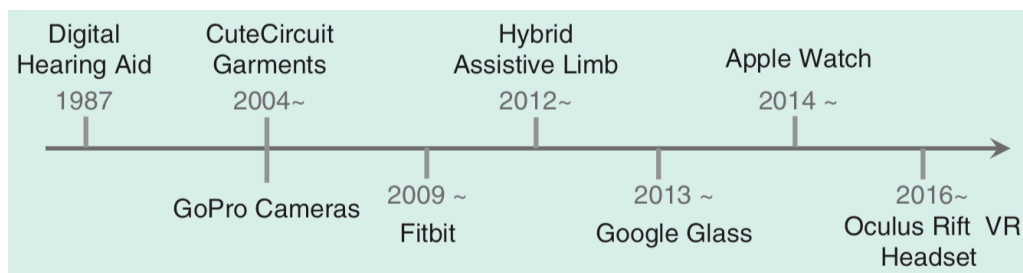


Figure 2.5: A time line of remarkable wearable devices.

The Fitbit Tracker, a small gizmo clipped onto clothing, was released in 2009. It can measure a wearer's movements around the clock and calculate his/her steps taken, sleep quality, and so forth for long-term tracking and analysis. The Hybrid Assistive Limb, first unveiled in 2012, is a recent leader in exoskeleton technology that is a kind of robotic system worn by both healthy and disabled people. With joints and links that correspond to those of the human body, an exoskeleton supports the limb movements of its wearer with increased strength and endurance. It can expand the physical capabilities of able-bodied people, serve as an assistive device for disabled people, and be used for rehabilitation.

Google Glass, released in 2013, attracted great media attention and is one of the most remarkable wearable devices in recent years. It is an optical head-mounted display in the shape of eyeglasses. It can deliver texts and notifications, be activated by voice commands, and take photos and record video with on-board cameras. Although the development of the smartwatch started as early as the 1970s, it only started gaining mass-market attention in the 2010s with the introduction of new models by Samsung, Sony, Apple, and Razer, among others. An essential feature of the new-generation smartwatches is their connection with other smart devices, especially smartphones. Connected smart devices can share resources and cooperate on tasks.

Another increasingly popular wearable technology is virtual reality (VR) gear. With the help of a VR headset (such as the Oculus Rift), a person can become immersed in an artificial world that is comparable to real life. People can see, hear, and even interact with objects in this virtual world.

2.6 On-Body Sensor Research

With the technological advances in circuit design, microelectromechanical systems, and signal processing, sensors are becoming smaller and smarter. Wearable sensors can now be used to measure many kinds of signals where one of the important components of health-monitoring systems is physiological sensors. Table 2.6 lists some physiological parameters commonly measured by wearable sensors, i.e.,

Table 2.2: Categories of wearable sensors and some sample research.

Signals	Technical Features	Positioning	Applications
ECG	A three-lead ECG sensor	Thorax	Vital sign monitoring
Respiration rate	Hall-effect respiration	Thorax	Emotion recognition
Respiration flow	A hot-film flow sensor	Upper body	Respiratory disease monitoring
Heart rate	Photoplethysmography (PPG) signals	Behind two ears	Hypertension control
Blood pressure	24-h monitoring of pulse transit time	Upper arm	Cardiovascular risk screening
Body temperature	Temperature sensor in armband	Armpit of smart vest	Vital signs monitoring
EDA	Electrodes attached to fingers	Middle fingers	Emotion recognition
Blood glucose	Amperometric electrochemical sensor	Implanted under skin	Continuous diabetic monitoring
EEG	Electrodes on viscoelastic earplugs	In ears	Sleep monitoring
Acceleration	Multiple triaxial accelerometers	Forearm	Respiratory diseases monitoring
Orientation	Triaxial accelerometer and gyroscope	Prosthetic socket	Gait analysis
Image	Smartphone camera	Waist	Fall detection
Throat sounds	Sounds separated from respiration	Over laryngopharynx	Ingestive behavior monitoring
Environmental sounds	Personal and social activities	In front of chest	Activity recognition
Strain	Jaw motions of piezoelectric strain	Below the outer ear	Ingestive behavior monitoring
Force	Triaxial force sensors	Shoe sole	Gait analysis
RF Signal	Motions cause different radio patterns	Embedded in clothes	Activity recognition

electrocardiogram (ECG), respiration rate, respiration flow, heart rate, blood pressure, body temperature, electrodermal activity (EDA), blood glucose, and electroencephalogram (EEG). Many chronic diseases, such as cardiovascular disorders, respiratory afflictions, hypertension, and diabetes, require long-term monitoring of vital physiological signals. Besides, many physiological signs are related to people's emotional state and thus are measured for emotion recognition.

A key aspect of the wearable health-monitoring is the use of programmable devices to enable in-node signal processing, simple data fusion, and even online classification. Furthermore, with the diffusion of personal handheld devices, these systems replaced desktop and laptops with mobile network coordinators, thus allowing mobile computing, a form of human-computer interaction by which a computer is expected to be transported during normal usage.

Inertial sensors include combinations of accelerometers, gyroscopes, and magnetometers. Such devices measure accelerations and orientations, so they are often used for motion analysis and activity monitoring, especially for gesture recognition, gait analysis, and fall detection. Inertial sensors are often combined with other sensors to achieve better task performance or to realize more complex functions. For instance, Zhang et al. used accelerometers and EMG sensors to recognize sign language. Chen et al. chose inertial sensors and a depth camera for

human action recognition. Cao et al. combined an accelerometer, and an airflow sensor for respiratory disease monitoring. Liu et al. joined inertial sensors and pressure sensors for intent pattern recognition of lower-limb motion.

Digital cameras can be embedded in smartwatches and many other consumer gadgets. These wearable cameras provide continuous recordings of a user's surroundings from the first-person perspective. Ozcan et al. attached a smartphone to a subject's waist. Using the on-board camera and accelerometer, they could reliably detect falls. Red Green Blue-Depth (RGB-D) sensors, such as the Microsoft Kinect, can provide range information that greatly facilitates target detection and recognition. Neto et al. mounted an RGB-D sensor on top of the wearer's head to perform face detection and recognition.

Many personal and social activities can produce characteristic sounds, so using wearable audio sensors for activity monitoring is a valid approach. Zhan and Kuroda used a wearable audio sensor mounted on the front of the chest to recognize environmental sounds (e.g. a vacuum cleaner, washing machine, hairdryer, and telephone ringing) and determined what kind of personal or social activity was taking place. Except for the previously mentioned sensors, other wearable sensory instruments can be adopted depending on different application requirements. For instance, a piezoelectric strain sensor was attached below the outer ear to capture the motion of the lower jaw to monitor chewing events. Wang et al. adopted wearable RFID tags for activity recognition. These tags and antennas were attached to specific body parts so that various activities resulted in different radio patterns, which could distinguish these activities.

2.7 Health Related Applications of On-Body Sensors

Wearable devices and systems are applied to many application scenarios, among which health care and fitness-related applications are the most popular. The increasing elderly population leads to rising healthcare demands. More efficient health-care strategies based on wearable technology are needed to realize autonomous, proactive, and predictive clinical services. For many age-related

chronic and degenerative diseases, such as cardiovascular disorders, Parkinson's disease, and diabetes, long-term home-based monitoring is quite essential (see Table 2.3).

The MyHeart project developed smart textiles and embedded ECG sensors in clothing so that patients wearing smart clothes could be comfortably monitored anywhere. The HeartToGo project developed a cellphone-based wearable platform for continuous ECG monitoring and recording. Gravina et al. focused on the detection of cardiac defense response (CDR). Frequent CDR activations can pose a health risk to patients and can eventually develop into severe psychophysical disorders.

Salarian et al. developed an ambulatory monitoring system for Parkinson's disease patients. This system consisted of three inertial sensors attached to the trunk and shins, and a portable data logger fixed to the waist. Sitting, standing, lying, and walking activities could be reliably detected and recorded. Bachlin et al. also used inertial sensors to measure patients' movements. Their system could automatically detect freezing-of-gait symptoms and provide audio feedback to the patients to help them resume walking.

Cao et al. proposed a wireless portable monitoring system for respiratory illnesses, incorporating a hot-film flow sensor, an oximetry sensor, and an accelerometer to measure the patients' respiratory airflow, and posture. These parameters could be transmitted via Bluetooth to a mobile phone or laptop for telemedicine. To continuously monitor asthma symptoms and environmental

Table 2.3: The application areas of on-body devices.

Areas	Typical Study	Highlights
Cardiovascular disease	MyHeart project	Smart textiles, ECG sensors embedded in clothing
Parkinson's disease	Salarian et al.	Activity recording and analysis with three inertial sensors
Respiratory disease	Cao et al.	Use hot-film flow sensor, oximetry sensor, and accelerometer
Diabetes	Medtronic's CGMS	Electrode needs to be inserted in tissue fluid
Rehabilitation	Pan et al.	Rehabilitation with inertial sensors and fingertip skin stretch
Activity/Fitness	Seshadri et al.	Study on how activity trackers influence a wearer's lifestyle
Entertainment	Chanel et al.	Measure player's emotional movements in a dancing game

substances that affect asthma, Dieffenderfer et al. developed a low-power wearable system consisting of a wristband, chest patch, and handheld spirometer. The system tracks physiological parameters (such as ECG, skin impedance, PPG, and lung function) and environmental parameters (such as ozone exposure, ambient temperature, and humidity).

There are many commercially available products to continuously monitor blood glucose, such as Medtronic's continuous glucose monitoring system and Guardian systems. A tiny electrode is inserted under the skin to measure glucose levels in tissue fluid. The measured information is then shown by a display device to remind the user whether the glucose level reaches certain limits.

Robot-aided rehabilitation, especially the rehabilitation exoskeleton, provides patients with a more effective and stable rehabilitation process and reduces therapists' workload. A powered elbow exoskeleton system for poststroke physical rehabilitation was developed by Vitiello et al. This system featured double-shelled links, a 4-degrees-of-freedom passive mechanism, and an antagonistic compliant actuation system, which enabled both robot-in-charge and patient-in-charge rehabilitation exercise. Pan et al. focused on balance rehabilitation of post-stroke patients. They used wearable inertial sensors to detect patients' body-sway angles during balance rehabilitation exercise and to provide fingertip skin stretch feedback to the patients through a portable sensory augmentation device.

We separate fitness-related applications from health care-related ones, focusing on those that engage people in physical activities or motivate them to adopt a healthier lifestyle. Activity/fitness trackers are devices or applications that monitor fitness-related metrics, such as distance walked, calories consumed, and sleep quality. Many smartwatches and smartphone apps can be regarded as fitness trackers. Research has shown that wearing fitness trackers can help diabetic and obese people make positive lifestyle changes. They become more aware of their inactivity and try to conduct a more active lifestyle.

Athletes and professional sports teams are often interested in monitoring movements and workloads to improve athletic performance. Wearable movement

sensors (such as pedometers, accelerometers, gyroscopes, and global positioning system devices) and physiological sensors (such as temperature monitors and heart rate sensors) enable real-time measurements of biomarkers and vitals during training and competitive sports. These parameters can be analyzed to design more effective training programs that maximize performance and minimize injury.

Wearable technology is offering new platforms and devices to make gaming more visually and physically engaging. VR headsets can now be used with computer games to offer 360° vision and sound, enable the wearers to interact with what they see, and make gaming more immersive.

Wearable gadgetry can make gaming more emotionally engaging by monitoring a player's emotional response and making adjustments accordingly. Chanel et al. tracked Tetris game players' physiological signals (EEG, EDA, body temperature, and respiratory rate), assessed their emotional state and then adapted the game difficulty to maintain players' engagement.

Using wearable sensors, a game interface can be more intuitive and intelligent. Textile pressure sensors and inertial sensors were embedded in game players' trousers, belt, and socks to measure their movements during a dancing game. Then each player's dancing image was reproduced as a virtual professional dancer on the game screen. Such a mixed-reality presentation can increase the enjoyment of a game and trigger a player's lasting interest.

Chapter 3

FOOD INTAKE MONITORING

Food intake analysis is a crucial step to develop an automated dietary monitoring system. Processing of eating sounds deliver important cues for food intake monitoring. Recent studies on the detection of eating activity generally utilize multi-modal data from multiple sensors with conventional feature engineering techniques. This chapter presents an on-body sensor system and related signal processing techniques to classify different types of food intake sounds. We then target to develop a methodology for the detection of ingestion sounds, namely swallowing and chewing, from the recorded food intake sounds during a meal.

A piezoelectric throat microphone is used to capture consumption sounds from the neck. For the classification problem, recorded signals are firstly segmented and decomposed using the empirical mode decomposition (EMD) analysis. EMD has been a widely implemented tool to analyze non-stationary and non-linear signals by decomposing data into a series of sub-band oscillations known as intrinsic mode functions (IMFs). For each decomposed IMF signal, time and frequency domain features are then computed to provide a multi-resolution representation of the signal.

Our second approach is related to the detection problem. Proposed methodology relies on feature learning in the frequency domain using a convolutional neural network (CNN). Spectrograms extracted from the recorded food intake sounds through a laryngeal throat microphone are then fed into the CNN architecture. This system has a potential for attaining high detection rates of the swallow and chew events with high sensitivity/specificity and delivers a food intake monitoring under daily life conditions in future studies.

3.1 Classification of Eating Sounds

Eating behavior and food consumption patterns contribute towards the development of dietary monitoring, mental eating disorders such as bulimia or anorexia nervosa [Harari et al., 2001] and other neurological impairments like dysphagia [Crary et al., 2013]. People suffering from obesity are also at a higher risk of blood pressure and other cardiovascular diseases [Van Gaal et al., 2006]. Automatic dietary monitoring can maintain an important tool for individuals to change their eating behavior.

Existing solutions of dietary monitoring such as food recalls, food-frequency questionnaires [Labonté et al., 2012] and self-report diaries [Lansky and Brownell, 1982] achieve low accuracy as people have a tendency to report biased consumption, which causes an inaccurate estimate of daily consume. Objective and automated methods of food intake classification were introduced as a potential solution to replace manual reporting procedures. These methods measure the periods of food intake with minimum participation so they provide deeper insight into eating behavior by improving the detection and classification accuracy.

Videofluoroscopy and electromyography (EMG) are considered as the benchmark methods in food classification studies. Videofluoroscopy requires detailed preparation with risks of aspiration and it must be performed in a clinic, while EMG is too invasive due to frequently used electrodes placed over the muscles of the neck. Thus, new techniques have been used in recent years to develop alternative solutions for dietary assessment.

Chewing and swallowing are regarded as the two main stages of food intake mechanism. After the food is taken to the mouth, chewing performs consecutive food degradation to shape a food bolus that can be swallowed. Several systems have identified chewing and swallowing acoustically. For example in [Sazonov and Fontana, 2012], acoustic data is acquired from multiple microphones coupled with strain gauge sensor from the ear canal. Wearing such a system is very unpleasant for users in terms of daily usage. In contrast, we propose to use a single necklace-like device to collect acoustic signals non-invasively, which is more

convenient and comfortable. In another work, eating-related arm gestures were detected by accelerometers and gyroscopes [Amft et al., 2005a]. Spoon, mug or hand usage associated with food intake were identified. However, subject-dependent eating style may limit the usefulness of this approach.

Piezoelectric sensors, which produce an output voltage corresponding with the mechanical stress, are used in many on-body sensing applications. Recently, they have been applied to problems in the medical domain, such as identifying individual heart rates and respiration [Klap and Shinar, 2013]. Very few works describe attempts to use piezoelectric sensors for monitoring food ingestion, with several exceptions [Ertekin et al., 1998], although the detection of swallowing disorders such as dysphagia was the primary objective of their work.

In this section, we use an oscillatory data investigation tool called the Empirical Mode Decomposition (EMD) introduced in [Huang et al., 1998]. It has been a novel technique for analysis of non-linear and non-stationary time-series and recently used in various signal processing problems [Mandic et al., 2013, Alam and Bhuiyan, 2013, Gudupudi et al., 2015]. The EMD decomposes any arbitrary time-series signal into a set of components known as intrinsic mode functions (IMFs). IMFs hold different frequency scales due to the dyadic filter bank structure of the EMD [Flandrin et al., 2004].

Traditional speech processing systems have been using the source-filter theory of linear prediction for many years as a principal representation. This theory simulates the speech signal as an output of the vocal tract filter excited by a dynamic source. Although the speech and ingestion signals have different characteristics and production mechanism, both of them have non-stationary nature. Methods using short time Fourier transform (STFT) on the basis of linear prediction theory can extract the spectral information over short time windows (10-30 ms) alongside locally calculated Fourier transform.

We consider the EMD for the analysis of food ingestion signals rather than the conventional Fourier-based or source-filter approaches. One disadvantage of applying STFT to food ingestion recordings is the "a-priori" assumption of a linear

and stationary model (i.e. sinusoidal amplitudes and constant frequencies are covering the whole signal). Therefore, STFT only provides fixed time and frequency resolution over short segments. In contrast to the STFT, EMD is an adaptive, data-driven and "a-posteriori" tool which does not need "a-priori" basis.

To solve the issues related to the fixed division of STFT, wavelet transform (WT) have been developed around the late 80's [Daubechies, 1988]. It can be considered as an alternative representation of Fourier-based procedure with an adjustable window from which to replace sinusoidal function to mother wavelet. However, due to the additional complexity and dependency of mother wavelet selection, this transformation has not become so popular for non-stationary signal analysis. Considering food intake recordings, one of the main factors is the shortness and non-stationarity of total data segments. Therefore, Fourier-based spectral examination methods such as STFT and WT encounter limitations in comparison with EMD.

A significant problem of any classification scheme is to tackle the increasing number of variables which may grow up the computational burden and mitigate the classification accuracy as well. The ultimate aim is to extract the optimum features from the signals that can perform the best discrimination between groups. Feature reduction is the technique of selecting a subset of relevant features for building robust classification models. By removing the most irrelevant and redundant features from the data, the classification problem becomes more feasible.

In this chapter, the minimum redundancy maximum relevance (mRMR) algorithm, which was originally proposed by Peng et al., is utilized to solve the feature selection problem of microarray gene expression data [Ding and Peng, 2005]. It tends to select a subset of features having the maximum correlation within a class (relevance) and the minimum correlation between themselves (redundancy). Relevance can be calculated by using the F-statistic (for continuous features) or mutual information (for discrete features) and redundancy can be calculated using the Pearson correlation coefficient or mutual information in a similar manner.

One of the goals in this thesis is to develop non-invasive monitoring and automated methods for the classification of food intake. A methodology based on EMD tool is proposed to classify food intake from epoch-divided sound signals. This work also explores the usefulness of EMD analysis for multi-class food intake classification along with the support vector machine (SVM) classifier. Monitoring of the food intake sound is conducted by a throat microphone sensor located over laryngopharynx. The mRMR feature selection process is utilized to determine the most relevant set of features on different IMFs representing the original sound.

The primary novelty of this study is the analysis of food intake sounds through the EMD and the investigation of most relevant features for food intake classification by means of the mRMR. Although a number of studies have been presented for the automatic food classification, to our knowledge no work has been done on EMD analysis for ingestion sounds. We propose a solution using a novel decomposition to accurately distinguish different food intake types.

3.1.1 Methodology

Block diagram of the proposed food intake classification system is given in Figure 3.1. Throat microphone sound recordings are firstly processed through temporal windows of length L_W with 50% overlaps. Feature extraction defines a set of features over all the IMFs and residual signal, where a reduced dimensional feature is then selected by the mRMR. Finally, we perform food intake classification for each data segment.

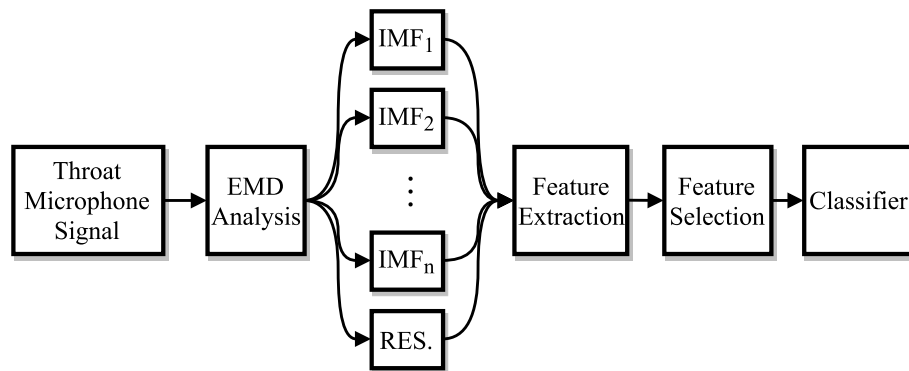


Figure 3.1: Block diagram of the proposed food intake classification system

EMD analysis

Throat microphones are skin-attached piezoelectric sensors that can record sound signals in the form of tissue vibrations. They can only capture low frequency bands, since tissue and bones act as a low-pass filter [Turan and Erzin, 2013, Turan and Erzin, 2016b]. In this thesis, we investigate EMD-based decomposition of the ingestion sounds, which are captured through the throat microphones.

EMD is a mathematical tool proposed by Huang [Huang et al., 1998]. It enables decomposition of non-linear and non-stationary data into dynamic basis functions known as IMFs. The decomposition is basically based on the simple assumption that any data consists of different sub-band modes of hierarchic oscillations. In principle, an IMF should satisfy the following two conditions:

1. The number of extrema points and zero crossings are either equal or differ at most by one.
2. The mean value of the envelopes defined by the local maxima and minima must be equal to zero.

Ultimately, the decomposition of a signal $s[n]$ as K number of IMFs and a residual r_K is given by

$$s[n] = \sum_{k=1}^K m_k[n] + r_K[n]. \quad (3.1)$$

The extraction of signal into different number of IMFs, $m_k[n]$, is known as sifting algorithm and summarized as the following:

1. Identify all local maxima and minima for a given discrete time signal $s[n]$.
2. Calculate the upper and lower envelopes, E_U and E_L , using cubic spline interpolation that connects the local extrema.
3. Find the mean value of the upper and lower envelopes, $E_\mu = (E_U + E_L)/2$ and subtract the mean from the original signal, $s[n] \triangleq s[n] - E_\mu$.

4. Execute the last three steps considering the updated $s[n]$ as the input signal until it satisfies the IMF conditions described above.
5. Subtract the first IMF, m_1 , from $s[n]$ to obtain the residue $r_1[n]$. This signal becomes $s[n]$ for the next iteration.
6. Iterate steps (1) to (5) on the residual $r_k; k = 2, \dots, K$ to find all the IMFs of the initial signal.

The sifting algorithm has two aims: 1) to eliminate riding waves, and 2) to make the wave signals more symmetric. Ideally after the first round of algorithm, the resulting signal at step (3) should satisfy the definition of an IMF. However, a small perturbed curve can be broaden to become local extrema which generates new coordinates although the fitting is accurate. Therefore, whole process is repeated as many times as is required to achieve an IMF at the final step of sifting algorithm. A detailed segment of the first iteration and obtained corresponding IMF is depicted in Figure 3.2.

This procedure can be terminated intuitively either when the residual becomes smaller than the predetermined threshold value or when the residual becomes a monotonic function where no more IMF can be extracted. Figure 3.3 illustrates the generation of the first four IMFs and the residual obtained from decomposing a sample almond segment of 500 ms length.

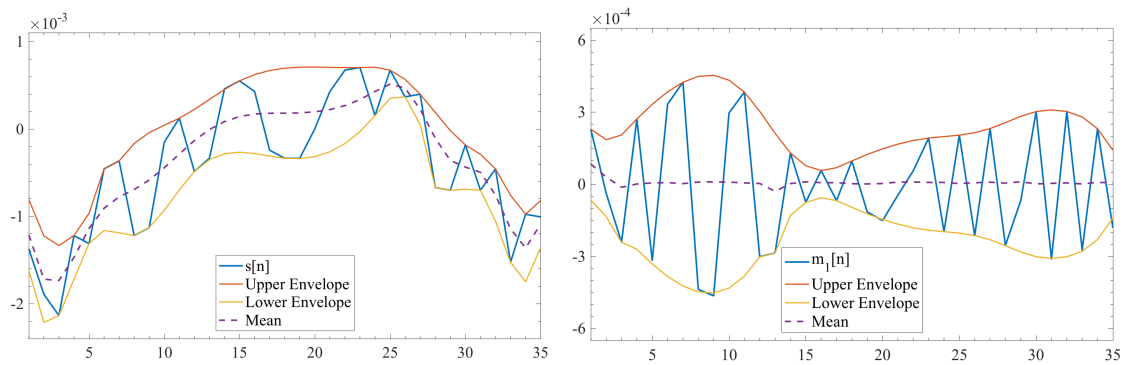


Figure 3.2: First iteration of sifting process (left), first IMF after repetition of the algorithm (right)

Clearly, the highest oscillation is embedded in the first IMF. This is an expected output of the sifting process which is based on diminishing scheme of the largest period component from data. Consequently, the components with the highest frequencies are stored in the first IMF. In other words, lower frequency content will be more dominant with the increasing order of the IMFs.

Accurate estimation of the IMFs may become an important factor because of the interpolation errors or an erroneous stopping criteria. Since the sifting process is vulnerable to the small changes and the choice of stopping criteria may produce different IMFs, the overall EMD procedure is not unique.

Finally, each IMF represents a simple oscillation as a counterpart to the harmonic function with variable amplitude and frequency as functions of time instead of constant amplitude and frequency. Furthermore, the oscillations should also be symmetric as compared to the local mean. The signal may have many different modes of oscillations at any time instance where each one is superposing on the others. In all implementations, we use the EMD toolbox available on Flandrin's web page.

Feature extraction

In this study, we pick 13 time domain and 3 frequency domain descriptors empirically to construct the feature vector for intake classification task. Note that extraction of these descriptors require minimal computational resources and they provide enough discrimination across different food intake events.

A summary of the feature set is given in Table 3.1. The first 11 parameters are low-level statistical descriptors of the signal. Zero-crossing rate and short-time energy are one of the frequently used descriptors in time-series analysis that reveal time-varying structure easily. The last three frequency domain features are extracted from the power spectral density estimate based on Welch's averaging.

Feature selection

The features, as defined in the previous subsection, are repeatedly extracted over all the IMFs and the residual signal to define an extended feature set $\mathcal{X}$. In feature selection, we rank the descriptors in the extended feature set using the mRMR selection algorithm, which targets to rank features with a high correlation within the class and a low correlation between themselves. The mRMR is a greedy algorithm that ranks one feature at an iteration incrementally and considers only pairwise redundancies between features. Therefore, features that are maximally dissimilar are selected to form a smaller feature vector with minimum redundancy.

A way of measuring the minimum redundancy among feature space $\mathcal{X}$ with n features x_i is,

$$\frac{1}{|\mathcal{X}|^2} \sum_{x_i, x_j \in \mathcal{X}} I(x_i, x_j) \quad (3.2)$$

where $I(x_i, x_j)$ denotes the mutual information of features x_i, x_j and $|\mathcal{X}|$ is the size

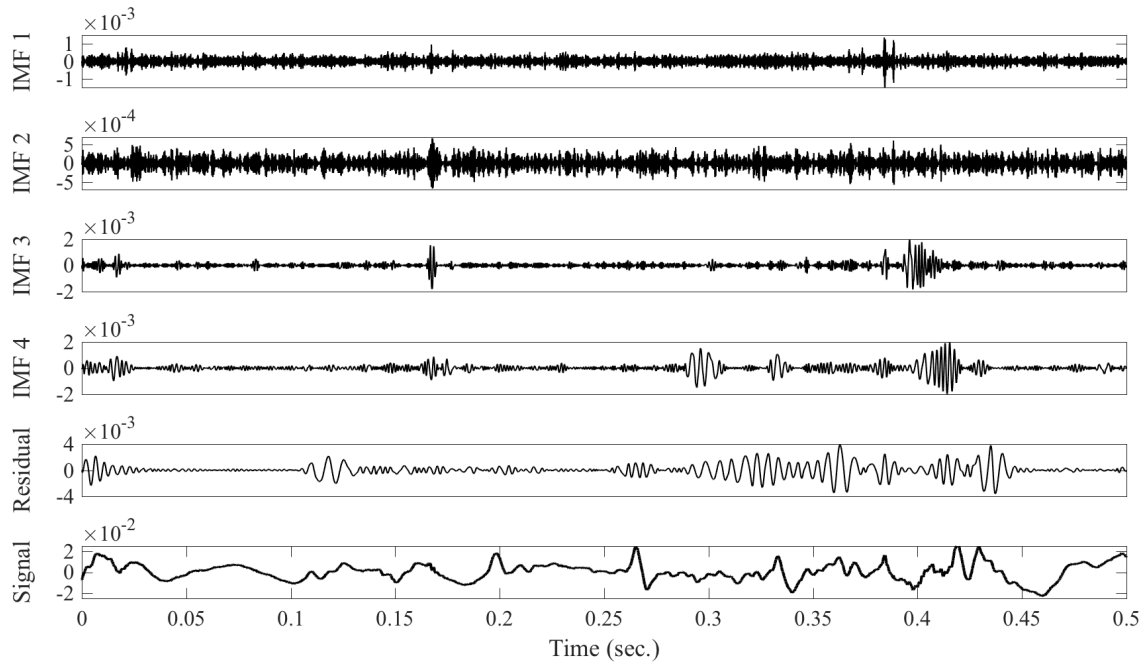


Figure 3.3: A sample empirical mode decomposition of a food intake signal

Table 3.1: Time and frequency domain descriptors in the feature set

#	Description	#	Description
1	Minimum Value	9	Third Quartile
2	Maximum Value	10	Interquartile Range
3	Mean Value	11	Range of Data
4	Standard Dev.	12	Windowed Energy
5	Kurtosis	13	Zero-Crossing Rate
6	Skewness	14	Maximum Power
7	First Quartile	15	Average Power
8	Median Value	16	Freq. at Max. Power

of feature space.

The mRMR criteria targets to minimize redundancy over accumulated mutual information $\sum_{i \neq j} I(x_i, x_j)$ while maximizing relevance of x_i within the target class c over the mutual information $I(x_i, c)$. Since an exhaustive search for the best subset is impractical, the selected set should be obtained in an incremental way. Suppose we already have $\mathcal{X}_{n-1}$, the feature set with $n - 1$ selected features. The task is to select n -th feature from the set $\mathcal{X} - \mathcal{X}_{n-1}$. This is done by selecting the n -th feature with the following maximization,

$$\arg \max_{x_j \in \mathcal{X}_{n-1}} \left[I(x_j, c) - \frac{1}{n-1} \sum_{x_i \in \mathcal{X}_{n-1}} I(x_i, x_j) \right]. \quad (3.3)$$

In this chapter, we adapt the implementation of mRMR defined in [Nguyen et al., 2014], which effectively solves the global optimization problem of (3.3).

Classifier

We utilized the SVM classifier for the food intake classification. SVM classifiers attain robustness and high generalization properties, which are useful when training group models with high variability between samples. It basically tries to find a

hyperplane that maximizes the separation of margins between data points given a weight vector ω and bias b . In this study, SVM classifiers are trained with a radial-basis function (RBF) kernel that forms $K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$. The width parameter γ is defined as $\gamma = 1/2\sigma^2$.

Our major interest is to reveal the discriminating capabilities of the proposed features through the EMD framework. The target here is much on the decomposition instead of the classifier type. Thus, one of the well-known classical techniques like SVM have been utilized.

3.1.2 Experimental Evaluations

Food intake dataset

We collect an in-house food intake sound dataset from a single subject using the IASUS NT3 throat microphone. The subject is a male and does not present any medical condition that would interfere with normal food intake. The IASUS NT3 throat microphone is placed over the trachea and sound recordings are captured at 16 kHz sampling rate. Time waveform and spectrogram of a sample recording with 7 sec duration are respectively given in the top and bottom plots in Figure 3.4.

The food items consumed during the intakes are selected to represent different physical properties of everyday foods such as "hardness" and "crunchiness". For example, cookie is a typical solid food and yogurt is a semi-solid food while almond is a hard crunchy food. The collected dataset includes a total of 481 recorded food intakes, which are summarized in Table 3.2.

mRMR ranking evaluations

The mRMR ranking of the features is applied both before and after the EMD processing for different experimental analysis. We limit EMD processing to extract up to four IMFs and a residual to reduce the computational cost. As suggested in [Ding and Peng, 2005], the mRMR-based feature selection performs better for quantized data as noise in continuous features could be minimized with discrete

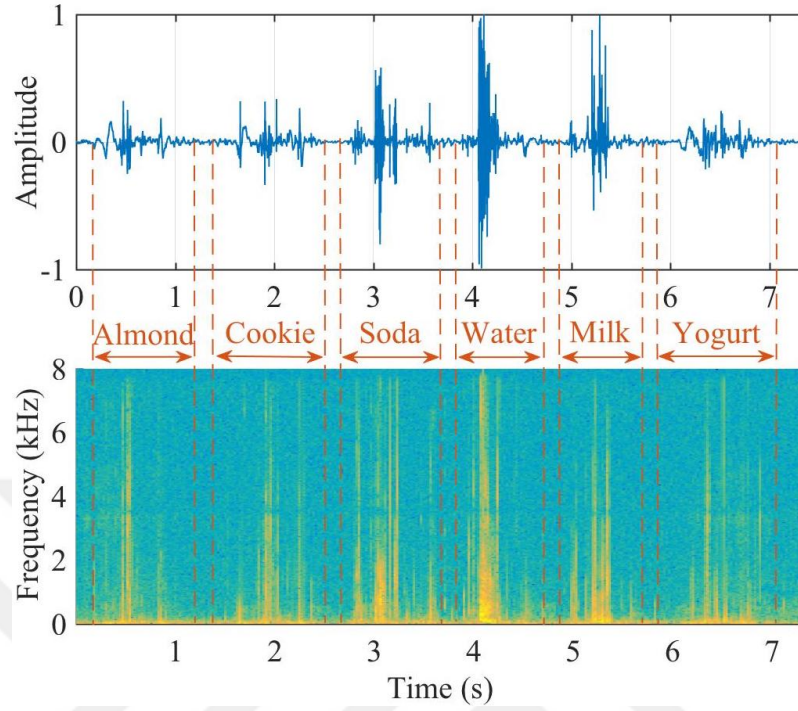


Figure 3.4: Waveform and spectrogram representations of different food intakes

variables. We also choose to quantize continuous valued features into 16 discrete centroids prior to mRMR ranking. In order to define mRMR-ranked feature

Table 3.2: A summary of the food intake dataset

Food Item	# of Intakes	Average Duration	Viscosity
Almond	84	$1.7 \pm 0.4s$	Solid
Cookie	65	$1.9 \pm 0.3s$	Solid
Soda Drink	71	$1.8 \pm 0.5s$	Liquid
Water	95	$1.5 \pm 0.3s$	Liquid
Milk	87	$1.7 \pm 0.4s$	Liquid
Yogurt	79	$1.6 \pm 0.5s$	Semi-Solid

representations, let us define all available feature space as

$$\mathcal{X} = \{x_{i,j}\} \text{ for } i = 1, \dots, 16, j = 0, \dots, 5 \quad (3.4)$$

where $\{x_{i,0}\}$ are the features extracted from the food intake sounds $s[n]$, $\{x_{i,5}\}$ are the features extracted from the residual in the EMD, and the features $\{x_{i,j}\}$ are from the first four IMF signals $m_j[n]$ for $j = 1, \dots, 4$.

We define four selected feature representations over the available feature space. The first feature set is the original food intake sounds, $\mathbf{F}_S = [x_{1,0} \cdots x_{16,0}]$. We then performed mRMR on this feature set and select the n most relevant features and denote them as $\mathbf{F}_S^n$. The last two feature sets are defined over the decomposed signals through EMD. First, we rank the descriptors in the feature set, where each descriptor is a 5 dimensional vector (4 IMFs plus a residual) in the decomposed representation, $\mathbf{x}_i^D = [x_{i,1} \cdots x_{i,5}]$. The third feature set is defined over these mRMR ranked descriptor vectors by selecting the n most relevant vectors as, $\mathbf{F}_D^{5n} = [\mathbf{x}_{1*}^D \cdots \mathbf{x}_{n*}^D]$, where $\mathbf{x}_{i*}^D$ is the i -th ranked descriptor vector. Finally, we perform mRMR on all components of the decomposed representation, $\{x_{1:5,1:16}\}$, and extract a ranked feature sequence, $\{x_{1*} \cdots x_{80*}\}$. The fourth feature set is defined as the n most relevant feature components over the EMD derived features as, $\mathbf{F}_{MD}^n = [x_{1*} \cdots x_{n*}]$.

Figure 3.5 presents ranking plots of the all feature descriptors over the EMD components. Feature descriptors are also ordered from top to bottom in their mRMR rankings as used in $\mathbf{F}_D^n$. Note that range of data, mean, frequency at maximum power and zero crossing rate are the top four descriptors in the mRMR ranking. On the other hand, low order IMFs are observed more relevant than the higher IMFs and the residual.

3.1.3 Results

We utilized four-fold cross validation to evaluate food intake classification performances. A temporal window length is chosen based on the following two

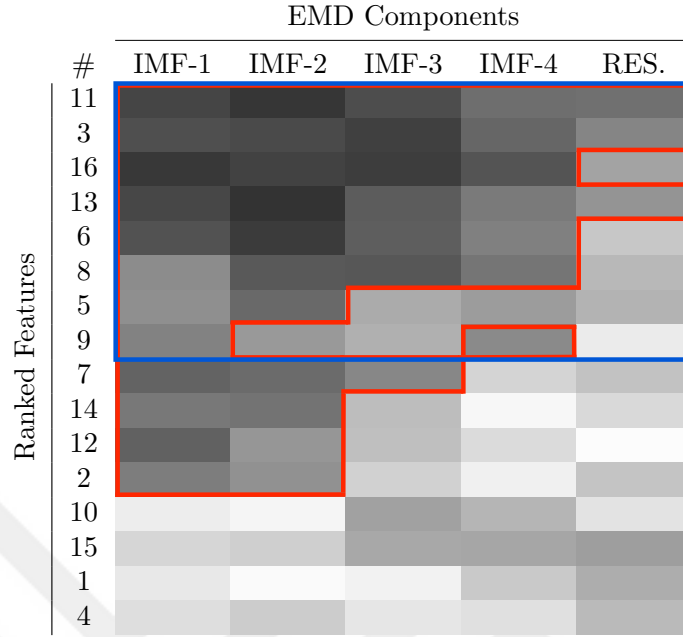


Figure 3.5: The mRMR ranking of the feature descriptors over the EMD components. Darker colors represent higher rankings.

criteria: 1) Portion length should preserve the non-stationarity behavior of the sensor signals, 2) There should be a fair comparison between classification accuracy and computational time required by the methodology. Based on this discussion, the intake data is processed and features are calculated over 500 ms overlapping segments with 50% overlap.

The classification accuracy is defined as the average of precision and recall rates. The precision and recall are defined as,

$$Precision = \frac{T_+}{T_+ + F_+}, \quad Recall = \frac{T_+}{T_+ + F_-}, \quad (3.5)$$

where the true positive (T_+) is the number of correctly classified intake windows, whereas the false positive (F_+) is the number of incorrectly classified windows and the false negative (F_-) is the number of times that classifier is failed to catch the successful intake events.

Table 3.3 shows a summary of the food intake classification accuracy results for the four selected feature sets as a function of the n most relevant features after

Table 3.3: Classification accuracies of the selected feature sets.

Accuracy as feature dimension n (%)				
n	4	8	12	16
$\mathbf{F}_S$				71.6
$\mathbf{F}_S^n$	66.3	69.8	72.5	71.6
n	20	40	60	80
$\mathbf{F}_D^n$	76.7	77.3	74.9	72.8
$\mathbf{F}_{MD}^n$	77.1	79.4	75.2	72.8

mRMR. The highest performance of each experiment is marked as bold for ease of visualization. The first feature set $\mathbf{F}_S$ is a baseline and sets a benchmark for the food intake classification. The mRMR-based selection over the baseline, $\mathbf{F}_S^n$, attains some classification improvement with the 12 most relevant descriptors and achieves 72.5% accuracy.

Performance of the mRMR ranked descriptor vectors $\mathbf{F}_D^n$ is on the third row of the Table 3.3. The $\mathbf{F}_D^n$ feature set achieves 77.3% accuracy with the 8 most relevant descriptor vectors for each component, which constructs a 40 dimensional representation (blue rectangular in Figure 3.5). The last row presents performance of the mRMR ranked features $\mathbf{F}_{MD}^n$, where the best classification accuracy is attained as 79.4% using the 40 most relevant features (red contour in Figure 3.5).

3.2 Eating Activity Detection

Healthcare applications became widespread after the extensive usage of wearable computation. Dietary monitoring is one of the most important utilization among other healthcare services. As a result of this, individuals should track what they eat to minimize the risk of these disorders.

Owing to the recent trend on mobile devices, reliable and robust monitoring

of eating activity can be performed through wearable sensors. The analysis of food intake behavior is a milestone to mitigate obesity and other eating-related disorders. The detection of chewing and swallowing events can be considered as a crucial step of this analysis. An algorithm of automated detection with high accuracy is therefore needed especially for a wearable solution.

A novel route in ingestion detection research is currently underway and is shaped by the sensing devices. Several sensor systems are investigated for food intake monitoring like EMG [Amft and Troster, 2006], EGG [Farooq et al., 2014], acoustic microphones [Walker and Bhatia, 2014] and image processing sensors [Kawano and Yanai, 2014]. Among them, acoustic-based modalities are commonly used with single- or multi-sensor configurations due to their non-invasive nature and ubiquitous accessibility. These sensors are used for various problems including food intake detection, food type classification, chew or swallow count and eating rate [Fontana et al., 2015]. However, acoustic modalities are quite vulnerable to environmental noise which results in degradation of performances under daily life conditions.

The deep neural network (DNN) is an emerging technique that has recently attained noticeable achievement in speech-related problems like recognition or synthesis [Zen and Sak, 2015]. Some DNN architectures have also demonstrated potential in speech understanding with computer vision tasks [Tür et al., 2012]. The requirement of a convolution operation in image recognition problem gives rise to a convolutional neural network (CNN) which has achieved strong outcomes similar to the DNN on speech recognition [Krizhevsky et al., 2012]. The convolutional structure followed by pooling in the CNN is a typical way to capture the class relations of objects in an image. Likewise, speech that is represented as a 2D-image or spectrogram over time and frequency, contains similar patterns in adjacent spectral bands. This fundamental resemblance between the vision and speech motivated us to analyze the patterns of eating activity obtained by the CNN systems for image classification.

Several research groups have already developed algorithms for detecting chew and

swallowing events but Amft et al. proposed one of the first studies that investigate detection problem [Amft et al., 2005b]. In this study, they placed microphones to the outer ear canal to capture the food intake sounds with a low signal-to-noise ratio. Then in [Amft and Tröster, 2008], they formed a multimodal detection system using k-nearest neighbor algorithm by unifying ear canal microphones with hand gestures captured from the accelerometer data and EMG sensor data located at the throat. Later, Amft introduced a wearable ear-pad sensor with a microphone inside of a headphone that can detect the chewing boundaries based on a feature similarity search algorithm [Amft, 2010].

Bone conducting microphones, placed in the outer ear, are also used for chewing related problems like period detection and counting of chew sessions by Shuzo et al. [Shuzo et al., 2010]. They capture the chewing events as maximums of the low-pass filtered sound signals. In another study, a Bluetooth headset is used for the detection of “crisp” and “non-crisp” foods by analyzing the signal-energy alterations based on Mel-frequency cepstral coefficient (MFCC) features [Nishimura and Kuroda, 2008].

In [Päßler et al., 2012], event detection is performed using hidden Markov models (HMMs). Both chewing and swallowing events are separated to build food intake units and the best fit of the spectral features is detected by the Viterbi algorithm. The same group is also developed several different algorithms for wearable implementation to carry on chew event detection and counting of chew strokes [Päßler and Fischer, 2014]. Among those algorithms, chewing detection based on signal energy maximums delivers the highest performance.

Spotting the swallowing sounds using microphones placed at the laryngopharyngeal region is proposed by Sazonov et al. [Sazonov et al., 2010]. Food intake periods of both solid and liquid item are detected by calculating the instantaneous swallowing frequency. Movements of the lower jaw recorded by strain sensors are also reported in [Makeyev et al., 2012]. In another study, a method based on auditory perception to detect the swallowing movement from laryngeal data is suggested in [Tanaka et al., 2012]. A detection system that

discriminates swallowing sounds from non-swallowing ones based on a moving average process is reported by Fukuike et al. [Fukuike et al., 2015].

A more recent study by Papapanagiotou et al. uses state-of-the-art learning methods to detect chew bouts [Papapanagiotou et al., 2017]. In a multi-sensor approach, a combination of the audio with photoplethysmography and accelerometer sensor readings are taken as input to CNN classifier architectures. This is the first attempt in the literature that introduces deep architectures to dietary monitoring research. However, it only achieves promising performances for isolated chew events with long duration input windows.

The investigations above-mentioned have various limitations. Multi-sensor system proposed by Amft et al. [Amft and Tröster, 2008] does not provide a comfortable wearable experience for the subject. In general, this is also valid for other studies that implement multi-sensor configurations as the wearable sensors are uncomfortable for daily usages. Other investigations like [Nishimura and Kuroda, 2008, Shuzo et al., 2010] focus on chew events or more specifically the detection of crisp-like foods. These works only interested in limited cases which are not applicable to the daily living conditions. On the other hand, ad-hoc algorithms subject to energy or maximum detection are using very primitive signal processing techniques [Päßler and Fischer, 2014]. Although these are simple enough to be used in wearable devices, again they suffer from generalization issue.

Presented here is an overview of the entry-level detection system of monitoring the daily ingestion activity using throat microphone signals. The primary goal of this section is to provide a detection framework that follows up the current advances in the learning paradigm. This new system does not depend on hand-selected features with traditional learning algorithms. Instead, it facilitates fine-tuned attributes with neural network structure to achieve superior performance.

3.2.1 Methodology

We target to detect food intake event classes from throat microphone recordings. Captured food intake sounds are translated into spectrogram representation using

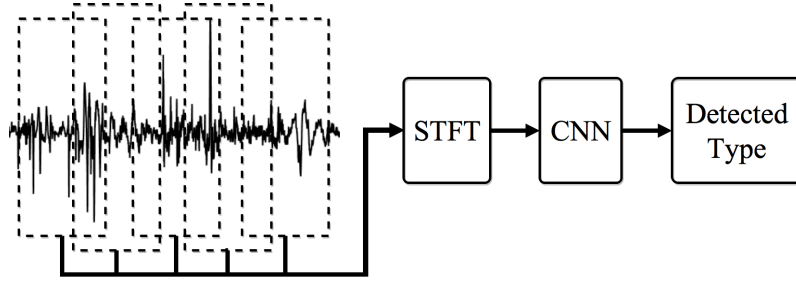


Figure 3.6: Block diagram of the proposed food intake detection system

short-time Fourier transform. Then, we utilize the CNN structure to learn features of time-frequency for the food intake classification problem. In this section, we first define our event detection system for the food intake events. Then, we define the spectrogram representation for the food intake sounds. Later, we describe structure of the CNN architecture for the targeted task.

Event Detection

The event detection system is based on a three-class classification model. Aside from chew and swallow events, a third class is defined as a *rest* event to represent all other activities like coughing, laughter, sneezing, speech, etc. Event detection is defined as a classification problem at time t of events chew, swallow and rest, over a temporal window centered at time t . The ground truth event label is taken as the event label at time t , that is the label at the center of the window. So, a slowly moving window over time can output a sequence of event classes, which are taken as output of the event detection system. Figure 3.6 illustrates the proposed food intake detection system.

Since the learning on whole clips is limiting the number of available examples for training, the spectrograms are partitioned to make intersection with each other. By doing so, we increase the number of images per class which motivates the CNN to end up with a better justification. The amount of shift is determined by the square resolution of spectrograms. Although there is no strong requirement about the resolution of images to be given as input to the CNN, square image size assists

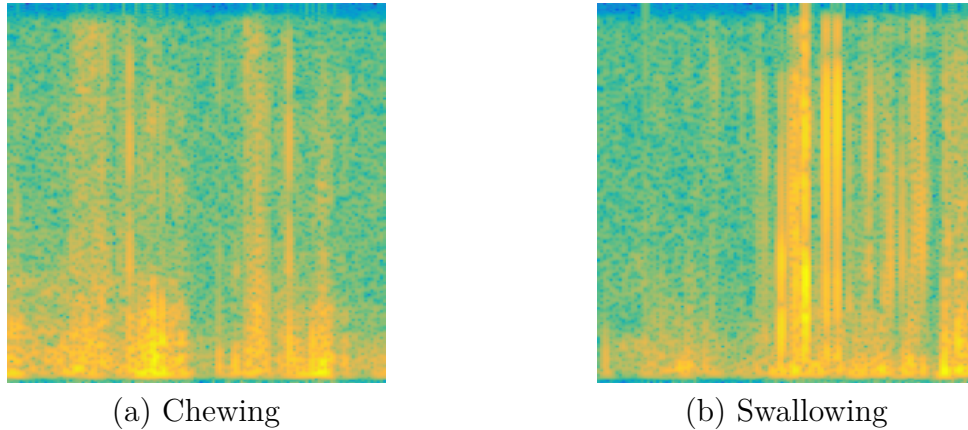


Figure 3.7: Sample spectrograms of food intake sounds

the network to reduce errors in matrix operations [Krizhevsky et al., 2012].

Spectrogram Representation

Spectrogram representation visualizes time-frequency energy distribution on a two-dimensional graph. The signal energy at a particular time and frequency is represented by the color-map intensity in which higher amplitudes are represented by brighter reddish colors. Spectrograms are used in time-frequency visualization of speech signals, and utilized in many speech processing problems, such as audio event classification, emotion detection and speech recognition [Yu et al., 2013, Mao et al., 2014]. They are extracted from the speech signal using the fast Fourier transform (FFT). In order to represent the temporal resolution, the signal is broken up into overlapping windows in the time-domain, and FFT transformed magnitude of the frequency spectrum for each window is calculated. This process generally corresponds to the squared log-magnitude calculation of the short-time Fourier transform (STFT) of the signal. Sample spectrograms of the chewing and swallowing activities are shown in Figure 3.7.

Convolutional Neural Networks (CNNs)

The CNN is considered as a modified version of the standard feed-forward neural network with some structural differences, which provide unsupervised feature

learning.

Instead of using fully connected hidden layers, a typical CNN model consists of a special network structure which are called as convolution and pooling layers. Visual data like spectrograms are represented as a set of feature maps after convolving the input with several filters. Each neuron is connected to a fixed-size portion of the spectrograms based on the position in their map. Therefore, the spatial layout of the data is preserved on the feature map and all neurons in a map share weights that yields lower number of trainable parameters compared to a fully connected layer.

A pooling procedure is utilized to each convolution feature map independently to reduce their resolution. Pooling layers subsample the feature maps which reduces data amount and inserts some translation invariance as well. The most common pooling operations are taking the max (winner takes all) or mean of the input cells. This down-sampling further improves invariance to translations. The spatial resolution of feature maps decrease at upper layers due to more convolution and pooling. In order to use the CNN for classification tasks, it should be wrapped up by one or more fully connected layers to combine the features with all other frequency bands. A soft-max output layer on covering the uppermost layer is then built to compute the posterior probabilities.

Traditionally, feed-forward neural networks uses logistic sigmoid or hyperbolic tangent as typical non-linear activation functions. However, recent implementations of deep architectures replaced them with alternative solutions which improves efficiency of the training process. One of the most common is the application of rectified linear units (ReLUs) which modifies the activation as thresholded at zero [Maas et al., 2013].

Deep neural architectures have a natural tendency to over-fitting. Even in convolutional neural networks, where the quantity of parameters is reduced through weight sharing, the number of estimated values is most of the times bigger than the number of training cases by an order of magnitude. This can result in poor out-of-sample generalization. One way to tackle this problem was introduced in the form of dropout learning [Hinton et al., 2012b]. The concept is quite simple,

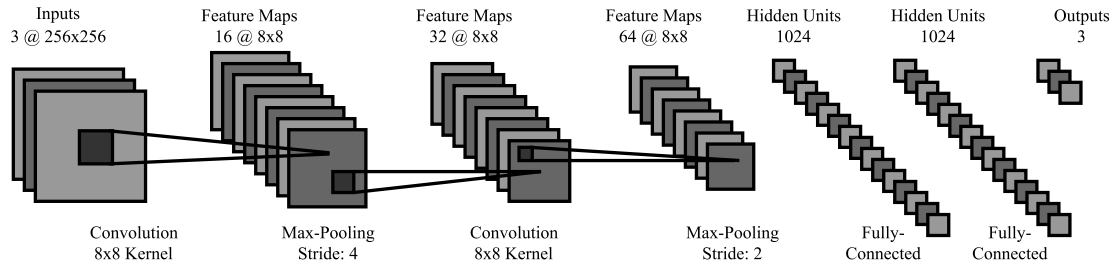


Figure 3.8: CNN architecture of the food intake detection system

yet highly effective. In each training iteration every hidden unit is randomly removed with a predefined probability, and the learning procedure continues normally.

3.2.2 Experimental Evaluations

Dataset

The data used in this chapter includes tracheal recordings of 8 subjects of 4 males and 4 females between 22 - 29 years old. Intake signals are collected in a laboratory environment with the iASUS NT3 throat microphone using the Sony IC recorder at 44100 Hz.

The iASUS NT3 throat microphone is placed over the neck. A total of 10 different tasks are recorded including chewing and swallowing of solids (potato chips, cake, biscuits, stick crackers, peanut and chocolate), swallowing of liquids (water, milk, fizzy drink, and fruit juice), as well as other non-related events (coughing, clearing the throat, speech or other kind of noise).

The food choices are selected to represent various food textures/categories including wet crispy, dry crispy, hard crunchy and soft, all of which impact the chewing and swallowing sounds. All participants have no history of chewing and swallowing abnormalities. Furthermore, specific chew and swallow events are manually annotated within each food intake cycle by visual and audio inspection



Figure 3.9: Sample images from the data collection sessions

of the recorded signals. Figure 3.9 shows sample images taken from a data collection event. Each subject visits the laboratory 5 times and consumes these 10 different food items at one recording session. Each session takes around 10 minutes and in total we have 276 minutes of intake recordings.

Data Preprocessing

From the tracheal activities of interest, swallowing has a bandwidth up to 1.5 kHz and chewing up to 3 kHz depending on the substance being chewed. Hence, we first down-sample the recordings to 8000 Hz sampling rate. Then, we apply a high-pass FIR filter with a cutoff frequency of 40 Hz to remove the low-frequency noise on the signal. Therefore, according to the Nyquist theorem, 8 kHz is sufficient to preserve important characteristics of chewing and swallowing but possibly not all of the other characteristics like coughing and clearing the throat [Larson et al., 2011].

The CNN structure is trained in a supervised learning manner by using a large food intake dataset with corresponding event labels. Event detection is performed for every 0.25 sec over a temporal window of size 1 sec. Hence, input spectrogram data is extracted over 1 sec temporal windows with a temporal shift of 0.25 sec. Within each temporal window, the spectrograms are computed over 30 msec frames with a stride of 1/256 sec to obtain square spectrogram images with 512 FFT sizes.

CNN Architecture and Training

Following the CNN architecture introduced in [Krizhevsky et al., 2012], our proposed CNN-based food intake detection system shown in Figure 3.8. Our network contains 2 convolutional layers, 2 max-pooling layers, and 2 fully connected layers with a soft-max layer. The input of network is a 256 x 256 spectrogram generated from throat microphone signals. The initial convolutional layer has 32 (8 x 8) kernels with a stride setting of 4 pixels. It is followed by a max pooling layer of size 8 x 8 with stride 4. Second convolutional layer has 64 kernels of size 8 x 8 and they are applied on the input with a stride 2 followed. The activations at two convolutional layers is also performed by ReLU. Then two fully-connected layers having 1024, 1024, and a soft-max layer with 3 neurons are appended individually. These outputs correspond to chewing, swallowing and rest event classes, respectively. Moreover, the first two fully-connected layers are followed by dropout layers having a ratio of 50% to avoid over-fitting.

More than 8000 spectrograms per each class are generated in MATLAB from all the files in our dataset. The training process is run for 30 epochs with a batch size set to 100. Initial learning rate was set to 0.01 with a decay of 0.1 after each 20 epochs. The training is performed on a single NVidia GeForce Titan XP GPU with 12 GB onboard memory.

Evaluation Procedure

In the 276 min food intake dataset, durations of the three event classes are not balanced, as well data has total silence segments. We observed that the least amount of samples are gathered from swallowing events, which takes around 1.5 sec on the average. In order to balance the event durations, after removing the silence segments, we kept around 33 min from each event class chew, swallow and rest. The resulting data from 8 subjects is 100 min and balanced across food intake event classes.

The proposed event detection system outputs event class estimates at 4 Hz. Over the stream of estimates, we calculate true positive, false positive, true negative and false negative rates. Based on these values, precision, recall, f-score and overall

Table 3.4: Performances of the proposed food intake detection

	Pre.	Rec.	Acc.	F1
CNN with LOSO	0.702	0.641	0.704	0.682
+ ReLU	0.747	0.656	0.708	0.691
+ Dropout	0.754	0.672	0.712	0.707
CNN with 2-fold	0.792	0.752	0.771	0.759
+ ReLU	0.803	0.753	0.782	0.763
+ Dropout	0.809	0.761	0.783	0.778

accuracy are calculated.

In the experimental evaluations, we utilized a subject independent and a subject dependent train/test protocol. In the first protocol, the training and testing is performed using leave-one-subject-out (LOSO) cross-validation to get the subject independent evaluations. This scheme corresponds to 8-fold cross validation, where in each fold food intake events of one subject are tested. In the second protocol, we utilized a 2-fold cross validation. In this setting, we split data into two equal subsets for each subject. Then, in the train/test protocol we guarantee that all subjects appear in training and testing within the 2-fold cross validation.

3.2.3 Results

In Table 3.4, we present evaluation results for each of the two evaluation protocols defined above. As expected, experiments with the 2-fold cross validation protocol yield higher performances than the LOSO protocol. Results show that the CNN with 2-fold cross validation protocol achieves 0.792 precision, 0.752 recall and 0.771 accuracy. In order to improve network performance, we apply a simple aggregation method that replaces activation function with ReLU. However, it does not add perceivable progress over the base CNN systems. Moreover, a dropout layer is also

used to prevent over-fitting, which delivers slight improvement for both experiments. With these modifications, the best accuracy results are attained as 0.712 and 0.783 for the subject independent and subject dependent settings, respectively. Performance degradation of the subject independent evaluations is limited, which indicates CNN is learning good cross-subject features for food intake detection.

Event class based evaluations are given as confusion matrices in Table 3.5 and Table 3.6 for the subject independent and subject dependent settings, respectively. Although we observe consistent improvements in all event classes for the subject dependent experiment, the most significant improvements appears in the chewing event class. This is mainly due to the cross-subject mismatch of chewing activity which yields false negatives for the LOSO evaluation.

Table 3.5: Confusion matrix for the LOSO experiment

		Predicted		
		Chewing	Rest	Swallowing
Actual	Chewing	53%	11%	36%
	Rest	3%	82%	15%
	Swallowing	2%	19%	79%

Table 3.6: Confusion matrix for the 2-fold experiment

		Predicted		
		Chewing	Rest	Swallowing
Actual	Chewing	64%	7%	29%
	Rest	2%	88%	10%
	Swallowing	1%	16%	83%

3.3 Discussion and Conclusion

In this chapter, a food intake classification system is firstly presented. Throat microphone recordings are used through the mRMR feature selection scheme over the EMD based signal decompositions. The employed sensor technology is non-invasive and not sensitive to acoustic noise.

EMD is a data-driven technique for non-linear and non-stationary signal analysis. The initial idea is to produce sub-band signals that represent local oscillations of the signal and to iterate the procedure on slower components hierarchically. After this decomposition method, a finite number of oscillatory components called as IMFs are extracted from the data. Unlike the Fourier-based tools which map signals onto a fixed basis set, EMD is a unique signal decomposition tool that makes no prior assumptions about the data.

A significant feature of the EMD analysis is its dependency on only the data adaptively and its capacity of finding appropriate basis functions that may bring out the important information in signal. Actually, different IMFs may have different physical meaning especially valid for practical applications so one of the potential future plan is to determine the existence of this meaning.

In the experimental evaluations we test classification performance of the mRMR feature selection scheme both before and after the EMD based signal decomposition. We observed significant performance improvements in the decomposed representation with mRMR feature selection. While the classification accuracy of the baseline system was 71.6%, it improved to 79.4% using feature selection in the decomposed representation. This is an encouraging improvement, which indicates that the EMD based multi-resolution analysis creates a richer set of features and the mRMR successfully selects a smaller set of relevant features among them.

The technique utilized here is a high-resolution method used as an alternative to Fourier-based analysis, which seem as the standard technique especially for speech signals. However, a comparative study considering other time-frequency tools such

as spectrogram or autoregressive models is needed in the study of dietary monitoring to better interpret the pros and cons of them. Aside from such a study, this research has shown that the EMD has a promising application in the analysis of intake classification despite of the fact that it doesn't provide a complete theory rather an empirical analysis.

One limitation of our evaluation is the small dataset size. In future work, we plan to collect free-living data with longer duration (e.g. entire day or more) recordings over an extended set of food types and subjects. It would be also of interest to find new feature descriptors which can improve the discrimination performance.

Recently, CNN architectures have been widely used for unsupervised feature learning in various problems, where for many problems they deliver improved representations than the classical feature representations. This is mainly due to problem specific feature learning capability of the CNNs over abundant data.

In this section, we also define a novel food intake detection problem from throat microphone recordings using the CNN. Our network configuration consists of two convolutional layers followed by three fully-connected layers with three class outputs for the chew, swallow and rest food intake event classes. This new methodology automatically conducts feature learning in the time-frequency domain over the spectrogram images and performs three-class detection via the soft-max classifier at outermost layer of the CNN.

ReLU and dropout layers are also integrated to improve the overall performance. We train the proposed food intake detection system on a dataset collected from 8 subjects in laboratory settings. Under subject dependent and independent settings, we obtain 0.783 accuracy / 0.778 F-score and 0.712 accuracy / 0.707 F-score, respectively.

Chapter 4

TRANSFER LEARNING APPROACH

An assumption of traditional machine learning methodologies is the training and testing data have similar statistical characteristics. However, in some real-world machine learning scenarios, this assumption does not hold. In such cases, knowledge transfer, if done successfully, would greatly improve the performance of learning by avoiding much expensive data labeling efforts. Transfer learning has been demonstrated to be effective for many applications as it exploits knowledge present in labeled training data from a source domain to enhance the performance in a target domain, which has little or no labeled data. Currently, most transfer learning methods assume the source and target domains consist of the same feature spaces which greatly limits their applications. This is because it may be difficult to collect auxiliary labeled data that shares the same feature space as the target domain.

Heterogeneous transfer learning methods have been developed to address such limitations. It is characterized by the source and target domains having different feature spaces, but may also be combined with other issues such as different data distributions and label spaces. These can present significant challenges, as one must develop a method to bridge the feature spaces, data distributions, and other gaps which may be present in these cross-domain learning tasks. A recent advance in deep learning shows that transfer learning becomes more effective with high-level abstract features learned by deep models, and the ‘transfer’ can be conducted not only between data distributions and data types but also between model structures. This chapter contributes an analysis of current methods designed for performing heterogeneous transfer learning tasks to provide an updated, centralized outlook into current methodologies.

4.1 Introduction

The field of data mining and machine learning has been widely and successfully used in many applications where patterns from past information (training data) can be extracted in order to predict future outcomes. Traditional machine learning is characterized by training data and testing data having the same input feature space and the same data distribution. When there is a difference in data distribution between the training data and test data, the results of a predictive learner can be degraded [Huang et al., 2007]. In certain scenarios, obtaining training data that matches the feature space and predicted data distribution characteristics of the test data can be difficult and expensive. Therefore, there is a need to create a high-performance learner for a target domain trained from a related source domain. This is the motivation for transfer learning.

Research on transfer learning has attracted more and more attention in different names: learning to learn, life-long learning, knowledge transfer, inductive transfer, multi-task learning, knowledge consolidation, context-sensitive learning, knowledge-based inductive bias, meta learning, and incremental/cumulative learning [Raina et al., 2007].

In this context, transfer learning aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task. Rather than learning all of the source and target tasks simultaneously, transfer learning cares most about the target task. The roles of the source and target tasks are no longer symmetric in transfer learning.

Figure 4.1 shows the difference between the learning processes of traditional and transfer learning techniques. As we can see, traditional machine learning techniques try to learn each task from scratch, while transfer learning techniques try to transfer the knowledge from some previous tasks to a target task when the latter has fewer high-quality training data.

The need for transfer learning occurs when there is a limited supply of target training data. This could be due to the data being rare and expensive to collect or

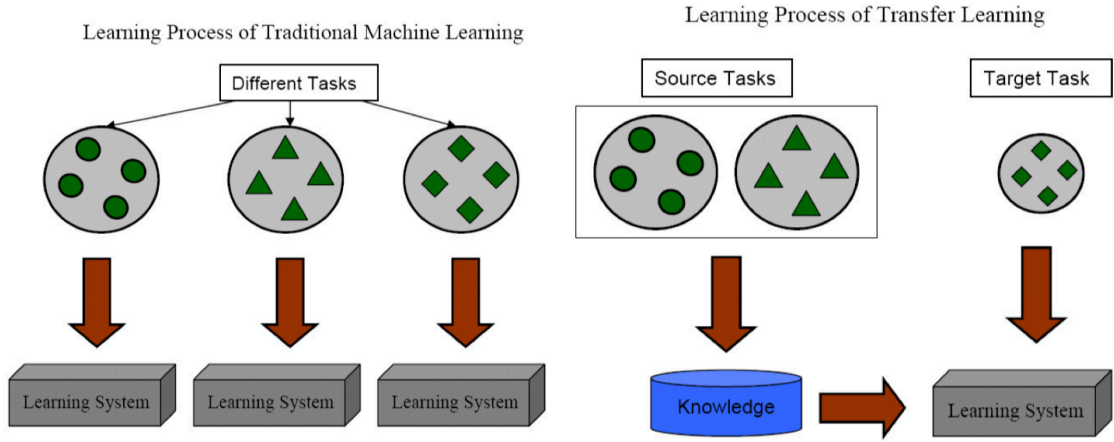


Figure 4.1: Different schemes between traditional learning (left) and transfer learning (right)

inaccessible. With big data repositories becoming more prevalent, using existing datasets that are related to, but not exactly the same as, a target domain of interest makes transfer learning solutions an attractive approach. There are many machine learning applications that transfer learning has been successfully applied to including text sentiment classification [Calais Guerra et al., 2011], image classification [Zhu et al., 2011], human activity classification [Cook et al., 2013], software defect classification [Ma et al., 2012], and multi-language text classification [Zhou et al., 2016].

4.2 Notations and Definitions

A domain $\mathcal{D}$ is defined by two parts, a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = x_1, \dots, x_n \in \mathcal{X}$. For a given domain $\mathcal{D}$, a task $\mathcal{T}$ is defined by two parts, a label space $\mathcal{Y}$, and a predictive function $f(\cdot)$, which is learned from the feature vector and label pairs x_i, y_i where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Now, $\mathcal{D}_S$ is defined as the source domain data where $\mathcal{D}_S = \{(x_{S_1}, y_{S_1}), \dots, (x_{S_n}, y_{S_n})\}$ and $x_{S_i} \in \mathcal{X}_S$ is the i -th data instance of $\mathcal{D}_S$ and $y_{S_i} \in \mathcal{Y}_S$ is the corresponding class label for x_{S_i} . In the same way, $\mathcal{D}_T$ is defined as the target domain data where $\mathcal{D}_T = \{(x_{T_1}, y_{T_1}), \dots, (x_{T_n}, y_{T_n})\}$ and $x_{T_i} \in \mathcal{X}_T$ is the i -th data instance of $\mathcal{D}_T$.

and $y_{T_i} \in \mathcal{Y}_T$ is the corresponding class label for x_{T_i} . Further, the source task is notated as $\mathcal{T}_S$, the target task as $\mathcal{T}_T$, the source predictive function as $f_S(\cdot)$, and the target predictive function as $f_T(\cdot)$.

Given a source domain $\mathcal{D}_S$ with a corresponding task $\mathcal{T}_S$ and a target domain $\mathcal{D}_T$ with a corresponding task $\mathcal{T}_T$, transfer learning is the process of improving the target predictive function $f_T(\cdot)$ by using the related information from $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$. From the definition of transfer learning, since $\mathcal{D}_S = \{\mathcal{X}_S, P(X_S)\}$ and $\mathcal{D}_T = \{\mathcal{X}_T, P(X_T)\}$, the condition where $\mathcal{D}_S \neq \mathcal{D}_T$ means that $\mathcal{X}_S \neq \mathcal{X}_T$ and/or $P(X_S) \neq P(X_T)$. The case where $\mathcal{X}_S \neq \mathcal{X}_T$ with respect to transfer learning is defined as heterogeneous transfer learning. On the other hand, the case where $\mathcal{X}_S = \mathcal{X}_T$ is defined as homogeneous transfer learning.

There are different strategies and implementations for solving a transfer learning problem. The majority of the homogeneous transfer learning solutions employ one of three general strategies which include trying to correct for the marginal distribution difference in the source, trying to correct for the conditional distribution difference in the source, or trying to correct both the marginal and conditional distribution differences in the source. The majority of the heterogeneous transfer learning solutions are focused on aligning the input spaces of the source and target domains with the assumption that the domain distributions are the same. If the domain distributions are not equal, then further domain adaptation steps are needed.

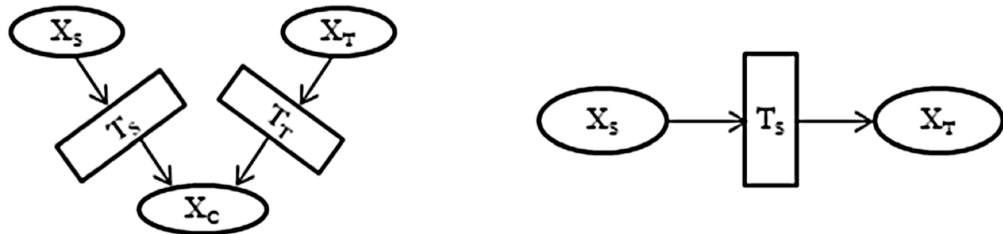


Figure 4.2: Different feature transformation schemes, the symmetric mapping of the source and target domains into a common latent feature space is on the left, the asymmetric transformation of the source domain to the target domain is on the right.

Inside these two strategies, transfer learning solutions fall into two categories when it comes to transforming the feature spaces. The first approach transforms the features of source through re-weighting to more closely match the target domain. This is referred to as asymmetric feature transformation and is depicted in Figure 4.2. The second approach discovers underlying meaningful structures between the domains to find a common latent feature space that has predictive qualities while reducing the marginal distribution between the domains. This is referred to as symmetric feature transformation.

4.2.1 Homogeneous Approach

The methodology of homogeneous transfer learning is directly applicable to a big data environment. As repositories of big data become more available, there is a desire to use this abundant resource for machine learning tasks, avoiding the timely and potentially costly collection of new data. If there is an available dataset that is drawn from a domain that is related to, but does not exactly match a target domain of interest, then homogeneous transfer learning can be used to build a predictive model for the target domain as long as the input feature space is the same.

In an early work, the feature augmentation method uses labeled source data and limited labeled target data [Daumé III, 2009]. In a transfer learning environment, there are scenarios where a feature in the source domain may have a different meaning in the target domain. This issue is referred to as context feature bias, which causes the conditional distributions between the source and target domains to be different. To resolve context feature bias, a method to augment the source and target space with three duplicate copies of the original feature set is proposed. More specifically, the three duplicate copies of the original set in the augmented source feature space represent a common set, a source-specific set, and a target specific set which is always set to zero. From the feature augmentation structure, a classifier learns the individual weights for the augmented feature set, which will help correct for any bias issues. The duplication of features creates feature separation between the domains and allows the final classifier to learn the

optimal feature weights. For the experiments, a number of different natural language processing applications are tested and in each case, the classification error rate is measured as the performance metric. However, when the source and target domains are very similar, this approach tends to underperform. The reason is duplication of feature sets represents irrelevant and noisy information when the source and target domains are very similar.

Multiple kernel learning allows for an optimal kernel function to be learned in a computationally efficient manner [Wu et al., 2011]. The paper by Duan et al. proposes to implement a multiple kernel learning framework for a transfer learning environment [Duan et al., 2012]. Instead of learning one kernel, multiple kernel learning assumes the kernel is comprised of a linear combination of multiple predefined base kernels. The final classifier and the kernel function are learned simultaneously which has the advantage of using labeled data during the kernel learning process. This is an improvement over [Pan et al., 2008] where a two-stage approach is used. The final classifier learning process minimizes the structural risk functional and the marginal distribution between domains using the maximum mean discrepancy measure [Borgwardt et al., 2006]. Pseudo labels are found for the unlabeled target data to take advantage of this information during the learning process. The pseudo labels are found as a weighted combination of base classifiers (one for each feature) trained from the labeled source data. A regularization term is added to the optimization problem to ensure the predicted values from the final target classifier and the base classifiers are similar for the unlabeled target data. Experiments are performed on the applications of video concept detection, text classification, and email spam detection. The methods tested against the feature replication method from [Daumé III, 2009] and outperformed it consistently.

The work by Long et al. is a joint domain adaptation solution that aims to simultaneously correct for the marginal and conditional distribution differences between the labeled source domain and the unlabeled target domain [Long et al., 2013]. To address the difference in marginal distribution between the domains, their proposed method requires a process to correct the conditional distribution

differences. Since the target data is unlabeled, pseudo labels (estimated target labels) are found by learning a classifier from the labeled source data. The maximum mean discrepancy distance measure is modified to measure the distance between the conditional distributions. Then, the features are used to train the final target classifier. Experiments are performed for the application of image recognition and classification accuracy is measured as the performance metric.

Another paper proposes an adaptation regularization based transfer learning framework for scenarios of labeled source data and unlabeled target data [Long et al., 2014]. This learning proposes to correct the difference in marginal distribution between the source and target domains, correct the difference in conditional distribution between the domains, and improve classification performance through a manifold regularization [Belkin et al., 2006]. The proposed framework will learn a classifier by simultaneously performing structural risk minimization, reducing the marginal and conditional distributions between the domains, and optimizing the manifold consistency of the marginal distribution. To resolve the conditional distribution differences, pseudo labels are found for the target data in the same way as proposed by Long et al [Long et al., 2013]. A difference between this approach and [Long et al., 2013] is adaptation regularization that learns the final classifier and minimizing the domain distribution differences simultaneously.

The paper by Pan et al. proposes a feature transformation approach for domain adaptation which does not require labeled target data [Pan et al., 2011]. The goal is to discover common latent features that have the same marginal distribution across the source and target domains while maintaining the intrinsic structure of the original domain data. The latent features are learned between the source and target domains in a reproducing kernel Hilbert space [Steinwart, 2001] using the maximum mean discrepancy as a marginal distribution measurement criteria. Once the latent features are found, traditional machine learning is used to train the final target classifier. This approach extends [Pan et al., 2008] by improving computational efficiency. Experiments are conducted for the application of WiFi localization where

the location of a particular device is being predicted. The source domain is comprised of data measured from different room and building topologies. The performance metric measured is the average error distance of the position of a device.

Another work proposes a spectral feature alignment that discovers a new feature representation for the source and target domain to resolve the marginal distribution differences [Pan et al., 2010]. It assumes an abundance of labeled source data and a limited amount of labeled target data. This approach identifies domain-specific and domain-independent features and uses the domain-independent features like a bridge to build a bipartite graph modeling relationship between the domain-independent and domain-specific features. If the graph shows two domain-specific features having connections to common domain-independent feature, then there is a higher chance the domain-specific features are aligned. A spectral clustering algorithm based on graph spectral theory [Chung and Graham, 1997] is used on the bipartite graph to align domain-specific features and domain-independent features into a set of clusters representing new features. These clusters are used to reduce the difference between domain-specific features in the source and the target domains. All the data instances are projected into this new feature space and a final target classifier is trained using the new feature representation. This method is well-suited for the application of text document classification where a bag-of-words model is used to define features. For this application, there are domain-independent words that will appear often in both domains and domain-specific words that will appear often only in a specific domain. This is referred to as frequency feature bias, which causes marginal distribution differences between the domains. However, this approach only addresses the issue of marginal distribution differences and does not address any context feature bias issues, which would represent conditional distribution differences.

Gong et al. proposes a domain adaptation technique called the geodesic flow kernel that finds a low-dimensional feature space for reducing the marginal distribution differences between the labeled source and unlabeled target domains [Gong et al., 2012]. Their kernel is constructed using the source and target input feature data, which projects a large number of subspaces that lie on the geodesic

flow curve. This curve represents incremental differences in geometric and statistical properties between the source and target domain spaces. A classifier is then learned from the geodesic flow kernel by selecting the features that are domain invariant. This work directly enhances Gopalan et al. [Gopalan et al., 2011] by eliminating tuning parameters and improving computational efficiency. Experiments are performed for the application of image classification where classification accuracy is measured as the performance metric. The tested measurements between the different source and target domain pairs have a high correlation to the actual test results, meaning the domains that are found to be more related to the higher classification accuracies.

Another method that equalizes marginal distribution of the labeled source and unlabeled target domains is proposed [Shi and Sha, 2012]. A discriminative clustering process is used to discover a common latent feature space that is domain invariant while simultaneously learning the final target classifier. The motivating assumptions for this solution are the data in both domains form well-defined clusters which correspond to unique class labels, and the clusters from the source domain are geometrically close to the target clusters if they share the same label. A one-stage solution is formulated to minimize the marginal distribution differences while minimizing the predicted classification error in the target domain. Experiments are performed for object recognition and sentiment classification where classification accuracy is measured as the performance metric. The approach described above is tested against a baseline approach taken from Weinberger et al. [Wei and Pal, 2011] with no transfer learning. Other transfer learning approaches tested include an approach from Pan et al. [Pan et al., 2011] and Gopalan et al. [Gopalan et al., 2011]. For the object recognition and text classification tests, the method proposed by Shi et al. [Shi and Sha, 2012] is best in comparison tests. An important point to note is the baseline method outperformed both [Pan et al., 2011] and [Gopalan et al., 2011] due to their two-stage domain adaptation processes. However, this paper offers a hypothesis that two-stage processes are actually causes negative transfer. In a typical two-stage scheme, the

first stage reduces the marginal distributions between the domains and the second stage trains a classifier with the adapted domain data. On the other hand, the hypothesis that the two-stage transfer learning process creates low performing learners does not agree with the results presented in the individual papers by [Pan et al., 2011] and [Gopalan et al., 2011] and other previously analyzed works.

4.2.2 Heterogeneous Approach

Heterogeneous transfer learning is the scenario where the source and target domains are represented in different feature spaces. There are many applications where this type of learning is beneficial. The applications that are covered in this section include image recognition, multilanguage text classification, single language text classification, drug efficacy classification, human activity classification, and software defect classification.

As is the case with homogeneous transfer learning solutions, whether the source and target domains contain labeled data drives the solution formulation for heterogeneous approaches. Data label availability is a function of the underlying application. The solutions surveyed in this paper have different labeled data requirements. For transfer learning to be feasible, the source and the target domains must be related in some way. Some heterogeneous solutions require an explicit mapping of the relationship or correspondence between the source and target domains. For example, the solutions defined for Prettenhofer et al. [Prettenhofer and Stein, 2010] addresses the heterogeneous scenario of a source domain containing labeled and unlabeled data, and a target domain containing unlabeled data. The structural correspondence learning technique from Blitzer et al. [Blitzer et al., 2006] is applied to this problem. This method depends on the manual definition of pivot functions that capture correspondence between the source and target domains. Effective pivot functions should use features that occur frequently in both domains and have good predictive qualities. Each pivot function is turned into a linear classifier using data from the source and target domains. From these pivot classifiers, correspondences between features are discovered and a

latent feature space is learned. The latent feature space is used to train the final target classifier. [Prettenhofer and Stein, 2010] uses this solution to solve the problem of text classification where the source is written in one language and the target is written in a different language. In this specific implementation, the pivot functions are defined by pairs of words, one from the target and one from the source, that represent direct word translations from one language to the other. The experiments are performed on the applications of document sentiment classification and document topic classification. English documents are used in the source and other language documents are used in the target. The baseline method used in this test trains a learner on the labeled source documents then translates the target documents to the source language and tests the translated version. An upper bound method is established by training a learner with the labeled target documents and testing with the target documents. The average results show the upper bound method performing the best and the method in [Prettenhofer and Stein, 2010] performing better than the baseline method. An issue with using structural correspondence learning is the difficulty in generalizing the pivot functions. For this solution, they need to be manually and uniquely defined for a specific application, which makes it very difficult to port to other applications.

Another work [Shi et al., 2010], referred to as heterogeneous spectral mapping, addresses the specific transfer learning scenario where the input feature space is different between the source and target $\mathcal{X}_S \neq \mathcal{X}_T$, the marginal distribution is different between the source and the target $P(X_S) \neq P(X_T)$ and the output space is different between the source and the target $\mathcal{Y}_S \neq \mathcal{Y}_T$. This solution uses labeled source data that is related to the target domain and limited labeled target data. The first step is to find a common latent input space between the source and target domains using a spectral mapping technique. The spectral mapping technique is modeled as an optimization objective that maintains the original structure of the data while minimizing the difference between the two domains. The next step is to apply a clustering based sample selection method to select related instances as new training data, which resolves the marginal distribution differences in the latent

input space. Finally, a Bayesian based method is used to find the relationship and resolve the differences in the output space. Experiments are performed for the applications of image classification and drug efficacy prediction.

The algorithm by Wang et al. [Wang and Mahadevan, 2011] proposes a manifold alignment process to perform a symmetric transformation of the domain input spaces. In this solution, there are multiple labeled source domains and a limited labeled target domain for a total of K domains where all K domains share the same output label space. This approach is to create a separate mapping function for each domain to transform the heterogeneous input space to a common latent input space while preserving the underlying structure of each domain. Each domain is modeled as a manifold. To create the latent input space, a larger matrix model is created that represents and captures the joint manifold union of all input domains. In this manifold model, each domain is represented by a Laplacian matrix that captures the closeness to other instances sharing the same label. The instances with the same labels are forced to be neighbors while separating the instances with different labels. A dimensionality reduction step is then performed through a generalized eigenvalue decomposition process to eliminate feature redundancy. The experiments are focused on the application of document text classification where classification accuracy is measured as the performance metric. The methods tested against include a canonical correlation analysis approach and a manifold regularization approach, which is considered the baseline method. The baseline method uses the limited labeled target domain data and does not use source domain information. The approach presented in this paper substantially outperforms the canonical correlation analysis and baseline approach; however, these approaches are not directly referenced so it is difficult to understand the significance of the test results. A unique aspect of this paper is the modeling of multiple source domains in a heterogeneous solution.

Another solution proposed by Qi et al. [Qi et al., 2011] addresses the application of image classification. The authors claim that image classification is inherently more difficult than text classification because image features are not

directly related to semantic concepts inherent in class labels. Image features are derived from pixel information, which is not semantically related to class labels, as opposed to word features that have semantic interpretability to class labels. Further, labeled image data is more scarce as compared to labeled text data. Therefore, a transfer learning environment for image classification is desired where an abundance of labeled text data (source) is used to enhance a learner trained on limited labeled image data (target). In this solution, text documents are identified by performing a web search on class labels. In order to perform the knowledge transfer from the text documents (source) to the image (target) domain, a co-occurrence matrix that can be programmatically built by crawling web pages is designed. Using the co-occurrence matrix, a common latent feature space is found between the text and image feature, which is then used to learn the final target classifier. Experiments are performed with the methods proposed Dai et al. [Dai et al., 2009], Zhu et al. [Zhu et al., 2011] and a baseline approach using a standard classifier trained on the limited labeled target data. The results show that [Zhu et al., 2011] is performing the best in 15% of the trials, [Dai et al., 2009] is the best in 10% of the trials, and the method in [Qi et al., 2011] is leading in 75% of the trials. As with the case of [Zhu et al., 2011], this method is very specific to the application of image classification and is difficult to port to other applications.

The work of Kulis et al. proposes an asymmetric transformation algorithm to resolve the heterogeneous feature space between domains [Kulis et al., 2011]. For this scenario, there is an abundance of labeled source data and limited labeled target data. An objective function is first defined for learning the transformation matrix. The objective function contains a regularizer term and a cost function term that is applied to each pair of cross-domain instances and the learned transformation matrix. The construction of the objective function is responsible for the domain invariant transformation process. The optimization of the objective function aims to minimize the regularizer and cost function terms. The transformation matrix is learned in non-linear Gaussian kernel space. This method is referred to as the asymmetric regularized cross-domain transformation.

In a similar study, Harel et al. uses limited labeled target data and multiple labeled sources where an asymmetric transformation is desired for each source to resolve the mismatch in feature space [Harel and Mannor, 2010]. The first step in the process is to normalize the features in the source and target domains, then group the instances by class in the source and target domains. For each class grouping, the features are mean adjusted to zero. Next, each individual source class group is paired with the corresponding target class group, and a singular value decomposition process is performed to find the specific transformation matrix for that class grouping. Once the transformation is performed, the features are mean shifted back reversing the previous step, and the final target classifier is trained using the transformed data. The experiments use data taken from wearable sensors for the application of activity classification. There are five different activities defined for the experiment which include walking, running, going upstairs, going downstairs, and lingering. The source domain contains similar (but different) sensor readings as compared to the target. The method proposed in [Harel and Mannor, 2010] is compared against a baseline method that trains a classifier with the limited labeled target data and an upper bound method that uses a significantly larger set of labeled target data to train a classifier. Their approach outperforms the baseline method in every test and falls short of the upper bound method in every test with respect to the balanced error rate.

The paper by Zhou et al. claims that previous heterogeneous solutions assume the instance correspondence between the source and target domains are statistically representative (distributions are equal), which may not always be the case [Zhou et al., 2014]. This method uses a labeled source, unlabeled target, and unlabeled correspondence data. When using correspondence data between source and target domains the method assumes that they are statistically representative (though this assumption may not always hold in real-world scenarios). The idea is to first learn an asymmetric transformation from the target to the source domain, which reduces the problem to a homogeneous domain adaptation issue. The next step is to discover a common latent feature space using the transformed data (from

the previous step) to reduce the distribution bias between the transformed unlabeled target domain and the labeled source domain. Finally, a classifier is trained using the common latent feature space from the labeled source data. This solution is realized using a deep learning method employing a marginalized stacked denoised autoencoder as proposed by Chen et al. to learn the asymmetric transformation and the mapping to a common latent feature space [Chen et al., 2012]. The steps of denoising autoencoder and heterogeneous feature mapping are recursively applied to the source and target data in each layer as to generate the different levels and feature transformations for the layers used in this deep learning approach. After such process, standard classification models can be used on the source data with the latent representation to build a target classifier. The experiments focused on multiple language text sentiment classification where English is used in the source and three other languages are separately used in the target.

Chapter 5

APPLICATIONS OF TRANSFER LEARNING

When faced with little or no labeled training data, a trained model will have an insufficient discriminatory ability and be unable to predict accurately. To handle this issue, transfer learning techniques have received significant focus in research communities. Specifically, heterogeneous transfer learning has been studied as it broadens the application of current transfer learning methods. There are two main approaches to solving the heterogeneous feature space difference. The first approach, referred to as symmetric transformation, separately transforms the source and target domains into a common latent feature space in an attempt to unify the input spaces of the domains. This chapter firstly analyzes a phoneme recognition problem using stacked denoising auto-encoders to transfer knowledge across different feature spaces and simultaneously correct the data bias on the transformed feature space.

The second approach referred to as asymmetric transformation that maps the source feature space to the target feature space to align the input features. This approach is best used when the class instances in the source and target can be transformed without context feature bias. With this assumption, there should be no significant distribution differences between the domains. Inside asymmetrical heterogeneous learning, a food intake classification problem is analyzed using a recently proposed framework known as teacher/student learning. Although this concept is proposed for a model compression framework, we modify teacher/student learning such that it can correct the distribution differences between the source and target domains. On the other hand, correcting the distribution differences can be problematic as the nature of a transfer learning environment that has minimal labeled target data. To compensate for this issue, we also organize the model to achieve better learning via an unlabeled parallel data.

5.1 Phoneme Recognition with Denoising Auto-Encoders

Reliable phoneme recognition is a vital prerequisite for an automatic speech recognition (ASR) training [Palaz et al., 2015]. A standard acoustic model of any ASR system firstly maps incoming speech segment to phoneme sequence. Then posterior probabilities match the most likely word to recognize spoken utterance. Hence, higher phoneme recognition rates yield better ASR performance.

In a conventional system, hidden Markov models (HMM) capture the temporal structure of speech signals, wherein each HMM state probability density function of short-time cepstral speech features is represented by the Gaussian mixture model (GMM). These GMM/HMM models have become less popular after the deep learning domination in speech recognition [Hinton et al., 2012a]. However, deep neural network (DNN) generally requires larger training data. GMM/HMM models, on the other hand, can be trained using relatively smaller data that makes them still attractive. DNNs may be used either as nonlinear feature extractors, where the features are used in a conventional GMM/HMM system [Sainath et al., 2012], or as so-called "hybrid" models where a neural network estimates class posterior probabilities replacing the GMMs [Mohamed et al., 2012].

Recently, convolutional neural networks (CNN) have been successfully utilized in image processing [Krizhevsky et al., 2012]. Compared to standard neural networks, the main difference is to process input in small localized parts. Since in the speech spectrogram formant structure of a phoneme may change practically from speaker to speaker, pooling the relevant local features makes CNN to be less sensitive of these variations. A recent hybrid model achieves state-of-the-art phoneme error rate (PER) on TIMIT corpus as 21.6% [Abdel-Hamid et al., 2014]. This framework uses a soft-max output layer to compute the posterior probabilities for HMM states. These posteriors are then used to estimate the likelihood of states per frame. Finally, likelihoods are sent to a Viterbi decoder to recognize the continuous stream of speech units.

Speech enhancement algorithms aim to improve the sound quality and

intelligibility. They are mainly used as a pre-processing step for speech communication and ASR. Statistical approaches such as minimum mean-square error (MMSE) filtering are primarily used in speech enhancement due to their low computational complexity. However, they do not work well in non-stationary degradation because of their Gaussian assumptions [Martin, 2005]. To overcome this weakness, we adopted a deep learning approach so as to learn a feature transformation which maps features to a latent space by reducing the bias in cross-domain correspondences.

Domain adaptation or transfer learning is a new machine learning paradigm, which targets to improve learning problem by transferring knowledge from source domain to target domain [Day and Khoshgoftaar, 2017]. In this section, we consider two low-quality speech sources: throat microphone (TM) recordings and narrow-band (NB) telephony speech. TMs are skin-attached piezoelectric sensors that can capture signals in the form of tissue vibrations. Although they are robust to environmental acoustic conditions, they can only capture low-frequency bands of speech signal since tissue and bones act as a low-pass filter. Furthermore, TM recordings are in general scarce. Similarly, the NB telephony speech has a pass-band in 300-3400 Hz, which reduces the listening effort. Although NB speech is not scarce, we take it as an alternative low-quality speech source in this part.

Transfer learning is also adapted to speech enhancement which aims to remove noise or reverberation. In a recent study, reverberations are considered largely as language-independent, thus an enhancement model trained with the data in one language has been applied directly to other languages [Xu et al., 2014]. Another recent study presents an adaptation of the speech enhancement generative adversarial network by fine-tuning the generator with small amounts of data [Pascual et al., 2018].

In an earlier work, a two-stage joint analysis framework was presented to improve TM-only speech recognition system [Erzin, 2009a]. First, a temporal correlation model is used to define the context, then the TM spectral features are mapped to enhanced spectral features through linear prediction.

Inspired by the transfer learning studies, this chapter investigates a generic model that can be used for any scarce speech signal source, which is suffering from quality and intelligibility. The proposed architecture combines several techniques that have already demonstrated success in other tasks [Glorot et al., 2011, Chen et al., 2012, Zhou et al., 2014]. It is based on the stacked denoising auto-encoders (SDAs), which can learn robust representations by recovering the original features corrupted by some form of noise or degradation.

Hence, the main contribution of this section is to define a transfer learning framework for feature augmentation from a source domain, where plenty of high-quality speech is available, to improve phoneme recognition for a low-quality target domain. This contribution is significant especially for improving speech recognition performance of low-quality and scarce speech signal sources, such as TM recordings, which can deliver robust speech recognition under adverse conditions. Furthermore, our SDA-based feature augmentation presents a novel contribution, which has not been studied before, to improve phoneme recognition in degraded channels.

5.1.1 Methodology

The proposed system, illustrated below, consists of four major modules which are described in the subsequent sections.

5.1.2 Feature Extraction and Augmentation

Mel-frequency spectral coefficients (MFSC) are extracted as the speech signal features. MFSC features are computed as the mel-frequency band energies of the speech spectrum as

$$\mathbf{x}_n(d) = \frac{1}{2\pi} \int \log |S_n(\omega)|^2 |\theta_d(\omega)|^2 d\omega, \quad d = 1, \dots, D \quad (5.1)$$

where S_n is the short-time Fourier transform (STFT) of the speech signal at frame n and θ_d is the triangular window centered at the d -th mel-frequency band.

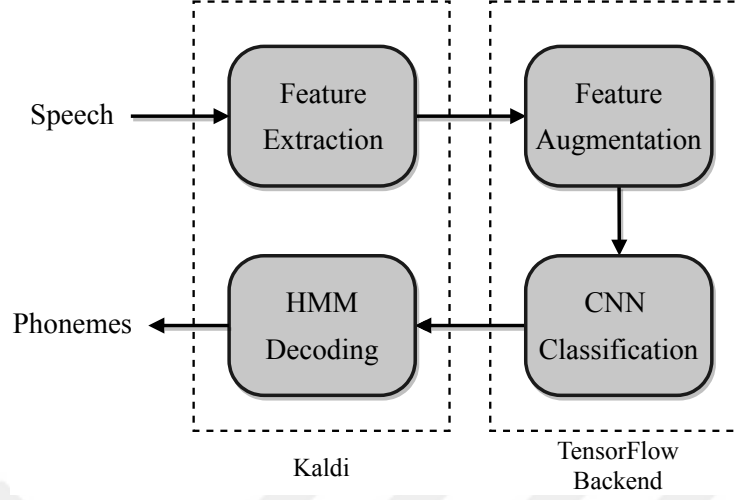


Figure 5.1: Block diagram of the proposed phoneme recognition system.

SDAs have been successfully used to learn new representations for feature transformation [Glorot et al., 2011]. A modified version of SDA called as marginalized SDA (mSDA) is proposed by Chen et al. [Chen et al., 2012]. In contrast to SDA, mSDA marginalizes noise and defines a closed-form solution. Therefore, it does not require stochastic gradient descent or other optimization algorithms to learn parameters. Furthermore, representations of the mSDA are as effective as the traditional SDAs, which are attaining almost identical accuracy in benchmark tasks [Zhou et al., 2014].

In this part, we adopt mSDA approach for learning MFSC latent space due to its simple single-level mapping structure. We follow the notation of Chen et al. and focus on the problem of MFSC enhancement throughout this section. We can set the target domain data as $\mathbf{D}_T = \{\mathbf{x}_{T_1}, \dots, \mathbf{x}_{T_M}\}$ and the source domain data as $\mathbf{D}_S = \{\mathbf{x}_{S_1}, \dots, \mathbf{x}_{S_N}\}$, where we have total number of M and N MFSC frames in the target and source domains, respectively. In our low-quality speech signal source model, we assume that we have access to the parallel source domain data that corresponds to the target domain data. Hence, the parallel data can be defined as $\mathbf{D}_P = [\mathbf{D}_T, \mathbf{D}_S^T]$, where $\mathbf{D}_S^T$ defines the parallel source domain data for the low-quality target data. Then mSDA learns weight matrices $\mathbf{W}_T$ and $\mathbf{W}_S$, and a mapping matrix

$\mathbf{G}$ to transform target and source domains to their latent representations.

We first learn the weight matrix $\mathbf{W}_T \in \mathbb{R}^{D \times D}$ by minimizing the squared loss function,

$$\mathcal{J}_1 = \sum_{i=1}^R \left\| \mathbf{D}_T - \mathbf{W}_T \mathbf{D}_T^{(i)} \right\|^2, \quad (5.2)$$

where $\mathbf{D}_T^{(i)}$ denotes the i -th corrupted version of $\mathbf{D}_T$. Solution of $\mathcal{J}_1$ depends on the random corruptions of input features. Then, the weight matrix can be expressed as a closed-form solution of ordinary least squares,

$$\mathbf{W}_T = \mathbf{P}_T \mathbf{Q}_T^{-1} \text{ with } \mathbf{Q}_T = \tilde{\mathbf{D}}_T \tilde{\mathbf{D}}_T^T \text{ and } \mathbf{P}_T = \overline{\mathbf{D}}_T \tilde{\mathbf{D}}_T^T \quad (5.3)$$

where $\overline{\mathbf{D}}_T$ is R -times repeated version of target data as $\overline{\mathbf{D}}_T = [\mathbf{D}_T, \dots, \mathbf{D}_T]$ and $\tilde{\mathbf{D}}_T$ is the corrupted version of $\overline{\mathbf{D}}_T$. In general, the larger R gives the more corruptions we average over and the less bias in estimation. Thus, by the weak law of large numbers, the matrices $\mathbf{P}_T$ and $\mathbf{Q}_T$ converge to their expected values as R becomes large. For these corruptions we prefer to use additive white Gaussian noise (AWGN) at different signal-to-noise ratios (SNR). Then the non-linearity is injected through an encoder function, which is learned together with the calculated reconstruction weights. A non-linear squashing function such as $\tanh(\cdot)$ is applied on the output of each mSDA to have a new feature representation as

$$\mathbf{H}_T = \tanh(\mathbf{W}_T \mathbf{D}_T). \quad (5.4)$$

Similarly, the weight matrix for the source, $\mathbf{W}_S \in \mathbb{R}^{D \times D}$, and $\mathbf{H}_S$ can be calculated.

We may cascade several mSDA layers by feeding the output of a layer as a new input to the next layer. Eventually at layer k , we have calculated pair of reconstruction weights $\mathbf{W}_{T,k}$ and $\mathbf{W}_{S,k}$ and their encoded higher-level representations $\mathbf{H}_{T,k}$ and $\mathbf{H}_{S,k}$. Let us denote $\mathbf{H}_{T,k}^P$ and $\mathbf{H}_{S,k}^P$ as encoded cross-domain features for the parallel target and source data, respectively. We then follow the feature mapping introduced by Zhou et al. [Zhou et al., 2014]

across encoded parallel instances $\mathbf{H}_{T,k}^P$ and $\mathbf{H}_{S,k}^P$. Here the main idea is to learn a feature transformation matrix at layer k , $\mathbf{G}_k \in \mathbb{R}^{D \times D}$ by minimizing the following objective,

$$\mathcal{J}_2 = \|\mathbf{H}_{S,k}^P - \mathbf{G}_k \mathbf{H}_{T,k}^P\|^2 + \lambda \|\mathbf{G}_k\|^2 \quad (5.5)$$

where $\lambda > 0$ is a regularization term on $\mathbf{G}_k$. The optimization problem of $\mathcal{J}_2$ has a closed-form solution and a summary of overall algorithm is tabulated below.

$$\mathbf{G}_k = [\mathbf{H}_{S,k}^P (\mathbf{H}_{T,k}^P)^\top] [\mathbf{H}_{T,k}^P (\mathbf{H}_{T,k}^P)^\top + \lambda \mathbf{I}]^{-1} \quad (5.6)$$

Algorithm 1: An overview of the feature augmentation

- Input** : Target and source domain data, $\mathbf{D}_T$ and $\mathbf{D}_S$
Parallel data of both domains $\mathbf{D}_T$, $\mathbf{D}_S^T$
Regularization parameter, λ
Total number of layers, K
- Output** : Transformed target data, $\mathbf{Z}_T$
Transformed source data, $\mathbf{Z}_S$
- 1 Start : Prepare $\mathbf{X}_T \triangleq [\mathbf{D}_T]$, $\mathbf{X}_S \triangleq [\mathbf{D}_S, \mathbf{D}_S^T]$ with initial values $\mathbf{H}_{T,1} = \mathbf{X}_T$ and $\mathbf{H}_{S,1} = \mathbf{X}_S$
 - 2 Use Eq. (5.6) and calculate $\mathbf{G}_1$
 - 3 **for** $k = 2, \dots, K$ **do**
 1. Apply mSDA for $\mathbf{H}_{T,(k-1)}$ and $\mathbf{H}_{S,(k-1)}$ and get pairs of $\mathbf{W}_T, \mathbf{H}_{T,k}$ and $\mathbf{W}_S, \mathbf{H}_{S,k}$.
 2. Use Eq. (5.6) and calculate $\mathbf{G}_k$
 - 4 **end**
 - 5 Compute the transformed feature instances
 $\mathbf{Z}_T = [(\mathbf{G}_1 \mathbf{H}_{T,1})^\top, (\mathbf{G}_2 \mathbf{H}_{T,2})^\top, \dots, (\mathbf{G}_K \mathbf{H}_{T,K})^\top]^\top$
 $\mathbf{Z}_S = [(\mathbf{H}_{S,1})^\top, (\mathbf{H}_{S,2})^\top, \dots, (\mathbf{H}_{S,K})^\top]^\top$
-

5.1.3 Convolutional Neural Network Model

In this section, we utilize a generic CNN-based classifier with the soft-max output. Consider a CNN input, $X_c = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_b, \dots, \mathbf{x}_B] \in \mathbb{R}^{B \times D}$, which is a sequence of B

consecutive frames of MFSC features, $\mathbf{x}_b \in \mathbb{R}^D$. Layer activations can be computed by a convolution-like operation of each filter followed by a non-linear activation. This is a standard CNN layer where weights and biases are shared among different layers. Alternatively, the weights may be partially shared [Abdel-Hamid et al., 2012], yet full weight sharing can also work sufficiently well and easy to implement.

After calculating local energy concentrations, CNN utilizes a max-pooling layer, which generates a low-resolution of the non-linear activations by taking the maximum activation within a pre-defined window. Finally, the top layers in CNN are fully-connected as in a feed-forward neural network. These layers combine the local structures extracted in the lower layers. In our hybrid model of phoneme recognition, posterior probabilities of HMM states are obtained using the top soft-max layer.

5.1.4 Hidden Markov Model Decoding

HMM is a widely used statistical model for temporal patterns. The main difference between GMM/HMM and CNN/HMM is the observation models. In GMM/HMM system, observation probabilities are modeled using GMMs under the maximum likelihood criterion. The context-dependent CNN/HMM systems use CNNs to classify the acoustic model into classes, which correspond to the HMM states. Therefore, the output of CNN predicts an estimate of the posterior probability, $\mathbf{P}(h_j|\mathbf{x}_n)$, of HMM state j given observation $\mathbf{x}_n$ at frame n . According to the Bayes' law, the emission probability $\mathbf{P}(\mathbf{x}_n|h_j)$, can be calculated as,

$$\mathbf{P}(\mathbf{x}_n|h_j) = \mathbf{P}(h_j|\mathbf{x}_n) \cdot \mathbf{P}(\mathbf{x}_n) / \mathbf{P}(h_j) \quad (5.7)$$

that can be simplified to $\mathbf{P}(\mathbf{x}_n|h_j) \propto \mathbf{P}(h_j|\mathbf{x}_n) / \mathbf{P}(h_j)$ where $\mathbf{P}(h_j)$ is the prior probability of state j calculated from the training data. In other words, we scale the CNN outputs by priors to get emission probabilities. For phoneme recognition task, we train HMMs for J discrete classes. For a new speech input $\mathbf{x}_n$, the corresponding phoneme class, $j^* = \arg \max_{1 \leq j \leq J} \mathbf{P}(\mathbf{x}_n|h_j)$ is calculated by the Viterbi

algorithm.

5.2 Knowledge Distillation in Food Intake Classification

Assistive devices and technologies in healthcare have been accelerating the development and integration of modern engineering approaches. A recent phenomenon, known as Smart and Connected Health (SCH), aims effective solutions in a wide variety of areas including pervasive computing, advanced analytics, and sensor integrations to facilitate major changes in healthcare practice.

The rapid emergence of obesity has triggered research for prevention and treatment. Today, 2.1 billion people (nearly one-third of the world's population) are either obese or overweight based on the data from 188 countries [Ng et al., 2014]. According to the World Health Organization, more than 3 million adults die each year because of obesity that makes it be one of the leading risks for the global deaths [Pozza and Isidori, 2018]. Moreover, a considerable amount of other factors such as diabetes or heart disease are associated with overweight and obesity. Thus, monitoring eating behavior is a crucial need for weight management, that can directly improve self-awareness and prevent unhealthy habits.

Many existing dietary monitoring tools rely on self-reports about what, how and when people eat. Food frequency questionnaires and 24-hour recalls are the most typical schemes, as they can after be considered as unreliable and error-prone methodologies [Schoeller et al., 1990]. These schemes are generally biased and depend on the user's ability to log their activities [Jacobs, 2012]. Underreporting and underestimating have therefore motivated the research towards systems that can automatically and objectively track intake activity [Amft et al., 2005b].

Automated dietary monitoring (ADM) can provide personal and detailed tracking on various aspects such as intake schedule, food amount and meal composition. Long-term monitoring would also provide valuable information about the diagnosis of chronic diseases in telemedical systems.

Latest advances in mobile and wearable computing have provided non-intrusive systems that support more reliable nutrient monitoring in recent years. These

gadgets have gained considerable popularity on health tracking [Fereidooni et al., 2017]. Previous approaches have been replaced by sophisticated systems that mitigate user burden and come up with real-time analysis of the collected data. Companies are now offering integrated dietary monitoring solutions for commercial wearables. According to a survey from International Data Corporation, the global wearables market during the first half of 2018 reached 27.9 million devices, up 5.5% from the previous year [IDC, 2018].

Although wearable devices and mobile applications present an opportunity to track eating behavior, they still require manual logging, such as taking a photo of the food or scanning a barcode [Patel et al., 2017]. There is, however, no available device that can automatically detect eating behavior in free-living conditions. Accessibility of such an ADM system would be indisputable assistance to personal wellness and healthcare research as well [Turan and Erzin, 2017].

In previous ADM research, there exist certain sensor configurations including condenser microphones and inertial sensors yet their monitoring capabilities are somehow controversial. This is because the traditional sensors capture not only the target signal but also the background noise. It has inspired the development of many non-acoustic vibration-based sensors to acquire signals more robustly [Erzin, 2009b]. Moreover, it is demonstrated that the piezoelectric design of a contact throat microphone was less susceptible to external noise than the condenser designs [Dekens et al., 2010]. These piezoelectric transducers are often used to capture tissue vibrations coming from speech production. Consequently, devices such as throat microphone (TM) or bone-conduction microphone (BCM) have been employed to provide viable signal sources for more robust and secure operations under both clean and noisy conditions. For a comprehensive review of these sensors, interested readers may look at here [Turan, 2013].

From the several alternatives of non-acoustic sensors, we employ a wearable system based on the TM. Essential reasons for this choice are social acceptability and cost efficiency of the TM modality. Since this sensor is positioned quite comfortably and steadily, it can be equipped with a power efficient source which is one of the

main targets in SCH [Chen et al., 2016].

A typical ADM system has several challenges: (i) identifying when and how long an individual performed food intake, (ii) identifying what and how much is consumed during the eating activity, and (iii) being user-friendly in real-world settings, i.e. it should be ubiquitous, robust to environmental noise, socially acceptable and easy to use. In this section, we focus on the accurate identification of which food is consumed. We take advantage of miniature, non-invasive transducers inside the TM sensor on the neck and aim at building a self-tracking system for an intelligent, unobtrusive device that keeps track of the consumed foods.

The two main problems of an ADM system, namely detection and classification of food intake, are partly solved by supervised learning approaches from manually annotated data [Turan and Erzin, 2017, Turan and Erzin, 2018]. However, non-acoustic modalities usually suffer from the lack of labeled data compared to the conventional acoustic sensors. This creates a bottleneck for current deep architectures, which are generally trained on massive amounts of data. As a special case, domain adaptation (or transfer learning) is an influential paradigm in which it adapts the information from one domain (known as source domain) that has a huge amount of labeled data to another related domain (known as target domain) that has little or no labeled data [Weiss et al., 2016]. The main addressed problem of this work is to define a transfer learning framework from the acoustic domain, where plenty of recordings are available, to improve intake classification of the TM recordings.

Currently, there is no accepted method for achieving high rates in the classification problem of an on-body ADM. One of the reason is that labeled data is scarce and manual labeling is expensive. This emphasizes the need for transfer learning in the ADM systems. In general, the domain adaptation problem arises when the source and target instances are generated from two different, but related distributions [Jiang and Zhai, 2007]. While fine-tuning of the parameters is a commonly used method for existing heterogeneous methods, applications of deep neural networks are limited in transfer learning studies [Sun and Saenko, 2016].

A possible solution is to benefit from teacher-student (T/S) learning, which has been shown persuasive performances in many problems [Hinton et al., 2015]. Inside the T/S learning, posteriors generated by the teacher model are used as soft labels to train the target domain student model. For our proposed framework, a student network with the TM recordings and a classifier with unlabeled data are jointly learned to minimize the Kullback-Leibler (KL) divergence between output distributions of the teacher and student models. In other words, the knowledge of posterior probabilities in the source domain is assessed by the teacher model where Hinton et al. refer to this as "distillation of knowledge" [Hinton et al., 2015].

In the original formulation, small student model tries to mimic the teacher that corresponds to obtaining a compressed model with comparable performance to the large teacher model [Hinton et al., 2015]. Unlike this scheme, our proposed method does not compress the model, but addresses student model to mimic teacher model as if the student network also performed intake classification. Thus, as opposed to the conventional way, our teacher and student models use different data as input, namely the TM data and the corresponding close-talk microphone (CM) data, respectively.

Recent advances in deep learning provide some useful insights on how to tackle a time-series classification like our problem. Specifically, recurrent neural network (RNN) and long short-term memory (LSTM) models include retention for temporal dependencies hence they have been proposed to solve the context in which they occur from sensor data [Ma et al., 2015, Hammerla et al., 2016]. The widely used LSTM classification (also known as fully-connected LSTM) provides a general framework for time-series data by training temporally concatenated LSTMs between the input sequence and the output label [Greff et al., 2017].

The fully-connected LSTM approach cannot efficiently capture the spatial correlation within the data especially like the case where spectrograms of the recorded signal are used as the input. Hence, in this part, we utilize our problem as a spatio-temporal sequence classification task in which both the input and classified target are spatio-temporal sequences. We adopt a convolutional LSTM (ConvLSTM) network which has convolutional structures in its transitions

[Xingjian et al., 2015]. Note that this new scheme differs from the combination of convolutional neural networks (CNNs) and LSTMs that first passes a long-term input into the CNN, then forwards the output into LSTM along with short-term features. Instead, ConvLSTM replaces the matrix multiplications underlying the internal structure of LSTM with convolution operations and have already demonstrated success in other tasks like video detection [Tang et al., 2018].

The main addressed question of this work is to overcome the limitations imposed by the lack of labeled data in the target domain that is generated by an on-body TM. Using the knowledge distillation, we propose a learning framework which delivers a noticeable gain in the ADM problem. In contrast to the standard formulation of knowledge distillation, our method provides a heterogeneous transfer learning structure in the distillation framework that reduces the distributional mismatch between cross-domain features.

The contributions of this section are threefold. First, we proposed a ConvLSTM topology that is capable of learning feature representations and modeling the temporal dependencies between food intake samples. Second, our learning scheme presents a novel framework for ADM via knowledge distillation, which has not been studied before. Here our aim is not a generalization of the complex model into a much lighter one instead we transfer the knowledge from a pre-trained network through unlabeled parallel data. Third, this framework performs intake classification using a single TM without giving importance to the recognition of individual chews or swallows. Previously proposed works commonly monitor the chewing or swallowing events by utilizing common machine learning methods. Yet this work combines several techniques related to neural network architectures which investigates a generic model towards end-to-end ADM systems.

Research in the area of intake monitoring dates back to 80s when the chews and swallows have been detected using oral sensors to measure the taste and satiety of foods [Stellar and Shrager, 1985]. Video-fluoroscopy and electromyography (EMG) are then considered as the gold standard of this research [Sato and Nakashima, 2006]. However, video-fluoroscopy depends on clinical equipment, while the EMG is

too invasive due to the electrode placement in muscles. Therefore, ongoing research is related to the use of non-invasive sensing modalities because they are lightweight for food monitoring that enables individuals to be free amongst different locations.

A collection of different strategies have been utilized for the complex problem of the ADM. Both selection and location of modalities are directly related to the output because a modality must be powerful in eating detection and be robust against adverse environments. Moreover, it needs to be acceptable in a wearable form in terms of portability, compactness and aesthetic pleasure.

Wearable accelerometers and gyroscopes have been one of the first modalities for detecting eating activities through on-body sensing. An approach for recognition of the eating gestures with a kinematic model of human forearm movements was proposed in [Zhang et al., 2009]. Papers that employ the accelerometers located on wrists usually focus on the differentiation between eating and drinking or activity recognition by estimating discrete actions such as taking chopsticks, stirring, or putting in mouth [Kim et al., 2012].

Acoustic modalities have been placed on the neck or ear regions. In an early study, Amft et al. placed microphones around the ear canal for classification of four food types [Amft et al., 2005b]. Their system was able to recognize individual chewing samples using a decision tree classifier. In a follow-up paper, the recognition of swallowing activity was investigated where they recorded signals from the surface EMG [Amft and Troster, 2006]. Although this study achieved decent recall, the precision remained low due to a large number of false positives.

In [Päßler et al., 2012], event detection was performed using hidden Markov models. Both chewing and swallowing events were separated to built food intake units and the best fit of spectral features was detected by the Viterbi algorithm. The same group was also developed several different algorithms for wearable implementation to carry on chew event detection and counting of chew strokes [Päßler and Fischer, 2014]. Among those algorithms, chewing detection based on signal energy maximums delivered the highest performance.

In another study, a Bluetooth headset was used for the detection of crisp and

non-crisp foods by analyzing the signal-energy alterations based on mel-frequency cepstral coefficients [Nishimura and Kuroda, 2008]. A similar paper used a microphone embedded in the Bluetooth headset and demonstrated the ability to detect eating among 12 different daily activities [Yatani and Truong, 2012].

In general, a primary drawback of acoustic systems is that the system performance can be significantly affected by environmental noise. To overcome this trouble, bone conducting microphones placed in the outer ear were also used for chewing related problems like period detection or counting of chew sessions [Shuzo et al., 2010]. It was used to capture the chewing events as maximums of the low-pass filtered sound signals. However, this modality is generally not acceptable due to its obstructive location for daily usage.

Neck-worn piezoelectric TM system was firstly used for real-time swallowing detection [Olubanjo and Ghovanloo, 2014]. Sazonov et al. also used the TM to spot the swallowing sounds [Sazonov et al., 2010]. Food intake periods of both solid and liquid items were detected by calculating the instantaneous swallowing frequency. In [Makeyev et al., 2012], a method based on auditory perception for the swallowing detection from laryngeal data was suggested. Another work presented a state-of-the-art learning method to detect chew bouts [Papapanagiotou et al., 2017]. The combination of audio with photoplethysmography and accelerometer sensor readings were taken as input to neural network classifier. This was the first attempt in the literature that introduces deep architectures to the ADM research. However, it only achieved favorable performances for isolated chew events.

Mainstream computational models in the ADM use ad-hoc algorithms subject to energy or maximum detection inside conventional learning techniques. Although these are simple enough to be used in wearable devices, they suffer from generalization issues. We recently proposed a convolutional neural network model for detection of both swallowing and chewing was recently proposed [Turan and Erzin, 2018]. Spectrograms extracted from the intake sounds through a single TM were fed into the network. This section was different from other works in the literature because it did not need multi-sensor configurations and attained

promising detection rates of the swallow and chew events.

5.2.1 *T/S Learning for Domain Adaptation*

Deep neural networks have recently exhibited state-of-the-art performances in many tasks such as speech recognition [Xiong et al., 2018], image classification [Zoph et al., 2018], and object detection [Wang et al., 2018]. However, top-notch systems usually require very wide and deep topologies, with numerous network parameters. Once learned, major drawbacks of such models are their storage and computational complexity requirements. For these reasons, wide and deep networks are not well suited for applications with memory or computation limitations.

More recently, knowledge distillation or T/S learning is introduced as a model compression framework, which simplifies the training using a teacher-student paradigm [Hinton et al., 2015]. Teacher-student paradigm describes a class of learning methods for training a smaller student network to mimic what a larger teacher network has been learned [Balan et al., 2015]. The teacher network typically is assumed to be trained previously. Original implementation penalizes the student according to a softened version of the teacher’s output. In other words, the student is trained to predict the output of the teacher, as well as the true classification labels. This strategy enables to compress the knowledge in an ensemble into a single model, which is lightweight and much easier to deploy.

Although T/S learning achieves improved performance for the student networks over several datasets, it has limitations such as difficulty with optimizing very deep networks [Yim et al., 2017]. To improve the performance of its training for deeper networks, Romero et al. proposed a hint-based training approach that uses the pre-trained teacher’s hint layer and student’s guided layer [Romero et al., 2014]. This allows the training of a student that is thinner than the teacher, using not only the outputs but also the intermediate representations learned by the teacher as hints to improve the training and final performance of the student.

There are several implementations that use knowledge distillation to compress acoustic models. Li et al. designed a larger deep model to train a smaller one by

utilizing the output distribution. In [Chan et al., 2015], Chan et al. transferred the knowledge from a recurrent neural network (RNN) model to a smaller deep network. A more recent study used an ensemble of individually-trained teachers whose predictions are then combined in a single model [Chebotar and Waters, 2016]. All of these methods utilize loss function to minimize the distribution differences between the two acoustic models. Their results show that these methods can compress acoustic models effectively with little performance loss.

Unlike the above-mentioned approaches, Li et al. evaluated the knowledge distillation without using transcribed data [Li et al., 2017]. This new formulation relies on parallel unlabeled data instead of the widely used distillation approach. The new formulation is implemented under two scenarios, adopting a clean acoustic model to noisy speech and an adult’s speech acoustic model to a children’s speech. A parallel corpus can be constructed from noisy replications of clean speech. Similarly, Yi et al. [Yi et al., 2018] proposed another knowledge distillation from the teacher model to the student model using parallel data to improve the performance of a far-field speech recognition task. They use the close-talking model to guide the training of the far-field model. In other words, their student model is trained to mimic the output of the teacher model by minimizing the loss between distributions.

Inspired by the above studies, we propose a new domain adaptation framework in a heterogeneous setup based on T/S learning paradigm. The teacher network is trained over abundant high-quality close-talk microphone recordings, whereas the student network takes TM samples as input and distills deep feature extraction capacity of the teacher over a parallel CM/TM dataset. Previous studies on knowledge distillation typically address homogeneous problems where the teacher and student domains lay in the same feature space. On the contrary, our proposed approach extends the general distillation framework to heterogeneous domain adaptation across the CM and TM domains. Using an unlabeled parallel data, we aim to bridge the gap in data distributions between the domains as experienced in a cross-domain heterogeneous transfer. This allows the use of a significantly larger

set of adaptation data and adds robustness to the resulting model.

In the following, we first present the vanilla version of the T/S learning to introduce the formalism of the T/S paradigm. Then, the proposed unsupervised heterogeneous domain adaptation framework is presented.

5.2.2 Vanilla Distillation

Neural networks typically employ class probabilities for a multi-class classification with "soft-max" output, which converts the predictions q_i (i.e. logits) for each class into a probability p_i , by comparing q_i with the other logits.

Let $\mathcal{X} = \{x_n \in \mathbf{R}^M \mid n = 1, \dots, N\}$ be a set of M dimensional input feature vectors and $\mathcal{Y} = \{y_n \in (1, \dots, C) \mid n = 1, \dots, N\}$ be the corresponding target labels, where N is the number of samples and C is the number of distinct categories.

For each input data and its label (x_n, y_n) , the standard cross-entropy (or KL divergence) criterion between a reference label and an arbitrary posterior distribution is,

$$\mathcal{L}_\theta(\mathcal{X}, \mathcal{Y}) = - \sum_{n=1}^N \sum_{j=1}^C \mathbf{1}\{y_n = j\} \log p(y_n = j | x_n; \theta) \quad (5.8)$$

where θ is defined as the model parameter and $\mathbf{1}\{\cdot\}$ is the indicator function. Here the reference label distribution is represented by a sparse one-hot representation. Instead of using the hard assignment labels, T/S learning uses posterior distribution of the label obtained from a teacher network with parameter Γ ,

$$\mathcal{L}_{\Gamma, \theta}(\mathcal{X}, \mathcal{Y}) = - \sum_{n=1}^N \sum_{j=1}^C p(y_n = j | x_n; \Gamma) \log p(y_n = j | x_n; \theta). \quad (5.9)$$

Contrary to the previous one-hot case, this new dense representation has many non-zero values, which leads to more powerful handling. Hinton calls this as "dark knowledge" which can be used to compressed student network with comparable performance to the teacher model [Hinton et al., 2015]. It enables the performance improvement of deep network models regardless of their underlying complexity.

5.2.3 Unsupervised Heterogeneous Domain Adaptation

Let us denote posterior distributions of the teacher and student networks as $p_T(\mathcal{Y}, \mathcal{X}^T)$ and $p_S(\mathcal{Y}, \mathcal{X}^S)$, where $\mathcal{X}^T$ and $\mathcal{X}^S$ stand for the teacher and student domain data, respectively. Note that in domain adaptation framework, teacher and students domains are referred respectively as source and target domains. However, we choose to use the teacher/student domain terminology to eliminate the confusion. Then, the KL divergence between these distributions can be

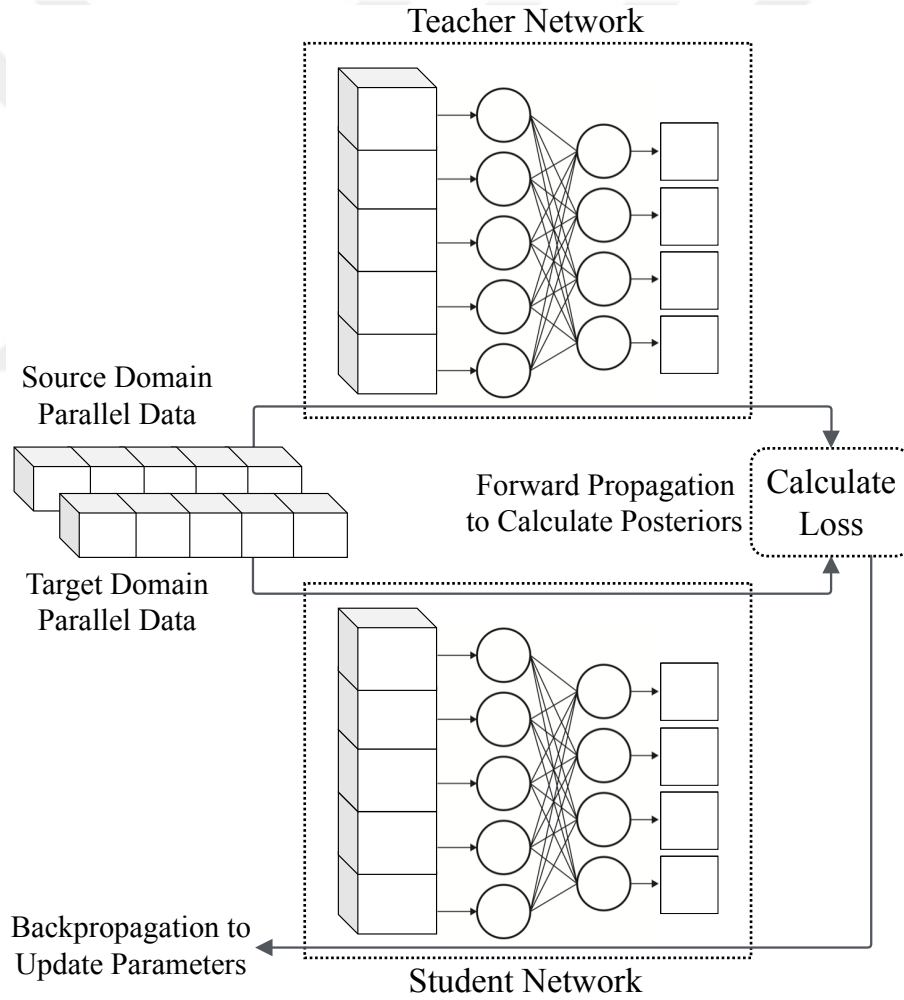


Figure 5.2: The architecture of T/S learning using parallel data for the proposed heterogeneous domain adaptation

rewritten as,

$$\mathcal{L}_{\Gamma, \theta}(\mathcal{X}^T, \mathcal{X}^S, \mathcal{Y}) = - \sum_{n=1}^N \sum_{j=1}^C p_T(y_n = j | x_n^T; \Gamma) \log \left(\frac{p_S(y_n = j | x_n^S; \theta)}{p_T(y_n = j | x_n^T; \Gamma)} \right).$$

Note that in this formulation teacher and student domains are heterogeneous, hence they are not necessarily instances of a single domain as in the vanilla T/S learning. The proposed domain adaptation framework, as shown in Figure 5.2, defines adaptation of the student network to distill feature extraction capacity of the teacher over an unlabeled parallel corpus.

In the proposed domain adaptation framework, the main target is to unify output distributions of the teacher and student networks for the paired samples of the parallel data. For this purpose, we can further simplify and re-write the cross-entropy loss function for each sample pair $(x_n^{T'}, x_n^{S'})$ in the parallel corpus as,

$$\mathcal{L}_{\Gamma, \theta}(\mathcal{X}^{T'}, \mathcal{X}^{S'}, \mathcal{Y}) = - \sum_{n=1}^N \sum_{j=1}^C p_T(y_n = j | x_n^{T'}; \Gamma) \log p_S(y_n = j | x_n^{S'}; \theta) \quad (5.10)$$

since $p_T(y_n = j | x_n^{T'}; \Gamma) \log p_T(y_n = j | x_n^{T'}; \theta)$ does not carry cross-domain information, it has little impact to the student network parameter optimization.

The algorithm 2 defines the proposed heterogeneous domain adaptation for the independent teacher and student domain data, $\mathcal{X}^T$ and $\mathcal{X}^S$, with the parallel domain data, $(\mathcal{X}^{T'}, \mathcal{X}^{S'})$.

Note that the proposed domain adaptation algorithm updates the student network using unlabelled parallel data. The supervision in the adaptation is obtained as the output of the teacher network within the parallel data. The performance of the adaptation is expected to improve with larger parallel data as well as with larger and richer teacher domain data, which shall provide better training for the teacher network.

5.3 Long-Short Term Memory for Sequence Modeling

In this section, we introduce the convolutional LSTM network architecture that is used within the proposed T/S learning based domain adaptation framework.

Given an input sequence $\mathcal{X}$, a standard recurrent neural network (RNN) computes the hidden vector sequence, $\mathbf{h} = \{h_1, \dots, h_N\}$ and output vector sequence $\mathcal{Y}$ by iterating the following equations,

$$h_n = \mathcal{F}(W_{xh} x_n + W_{hh} h_{n-1} + b_h), \quad (5.11)$$

$$y_n = \tanh(W_{hy} h_n + b_y), \quad (5.12)$$

where the W denotes weight matrices, the b denotes bias vectors and $\mathcal{F}$ is the hidden layer function, which is usually an element-wise multiplication of a sigmoid function. However, Long Short-Term Memory (LSTM) architectures are preferred over RNN, since they are better at finding and exploiting a long range context [Weninger et al., 2015]. Furthermore, LSTM prevents vanishing gradient, which is a critical problem for the standard RNNs [Pascanu et al., 2013].

In this study, we employ the ConvLSTM network, proposed by Shi et al. [Xingjian et al., 2015], which is a powerful model when the sequential data have spatial correlations. It has convolutional structures in both the input-to-state and state-to-state transitions. Hence, we are able to build a network model for

Algorithm 2: Heterogeneous Domain Adaptation

- 1 **Initialization:** Train the teacher and student networks separately with their independent domain data, $\mathcal{X}^T$ and $\mathcal{X}^S$, via the standard KL divergence criterion defined in (5.8);
 - 2 **repeat**
 - 3 For each mini-batch:
 - 4 - Do forward propagation of the teacher and student network using $(x^{T'}, x^{S'})$ to calculate $p_T(y|x^{T'}; \Gamma)$ and $p_S(y|x^{S'}; \theta)$;
 - 5 - Calculate KL divergence using (5.10), then do back propagation for the student network;
 - 6 **until** *convergence*;
-

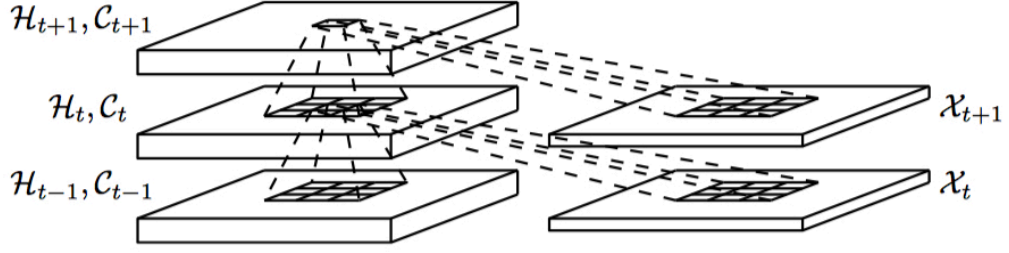


Figure 5.3: Inner structure of the ConvLSTM [Xingjian et al., 2015]

spatio-temporal sequence classification by stacking multiple ConvLSTM layers. As shown in Figure 5.3, each state transition of a ConvLSTM cell uses small portions of the input and the hidden states, which helps to improve efficiency on both the training and computations.

Similar to LSTM, ConvLSTM consists of an input gate i_t , a forget gate f_t , an output gate o_t and a memory cell c_t . However, all the inputs $\mathcal{X}_1, \dots, \mathcal{X}_t$, hidden states $\mathcal{H}_1, \dots, \mathcal{H}_t$ and gates are arranged as 3D tensors whose last two dimensions are spatial (rows and columns). The key equations of the ConvLSTM are,

$$\begin{aligned}
 i_t &= \sigma(W_{xi} * \mathcal{X}_t + W_{hi} * \mathcal{H}_{t-1}), \\
 f_t &= \sigma(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1}), \\
 o_t &= \sigma(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1}), \\
 c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc} * \mathcal{X}_t + W_{hc} * \mathcal{H}_{t-1}), \\
 \mathcal{H}_t &= o_t \circ \tanh(c_t),
 \end{aligned} \tag{5.13}$$

where "*" denotes the convolution operator and "o" denotes the Hadamard product. For simplicity, all the bias terms are omitted.

The ConvLSTM can learn the recent temporal history, since it accumulates past information in its memory cells. On the other hand, especially in speech like signals, information from both the forward and backward frames are important and correlative. Following to [Graves et al., 2013], we adopt a bidirectional ConvLSTM (Bi-ConvLSTM), which learns long-term dependencies between time

steps of time-series or sequence data. This can be useful when the network learns from the complete time-series at each time step.

In bidirectional recurrent networks, replacing each hidden sequence with the forward and backward sequences ensures the input from both directions at the level below. Thus, Bi-ConvLSTM captures temporal characteristics in both directions as depicted in Figure 5.4. The output with bidirectional information is defined as,

$$\mathcal{Y}_t = \tanh(W_{hy}^f * \mathcal{H}_t^f + W_{hy}^b * \mathcal{H}_{t-1}^b), \quad (5.14)$$

where $\mathcal{H}^f$ and $\mathcal{H}^b$ indicates the hidden states from forward and backward ConvLSTM units. Specifically, the backward ConvLSTM module can be specified as,

$$\begin{aligned} i_t^b &= \sigma(W_{xi}^f * \mathcal{H}_t^f + W_{hi}^b * \mathcal{H}_{t+1}^b), \\ f_t^b &= \sigma(W_{xf}^f * \mathcal{H}_t^f + W_{hf}^b * \mathcal{H}_{t+1}^b), \\ o_t^b &= \sigma(W_{xo}^f * \mathcal{H}_t^f + W_{ho}^b * \mathcal{H}_{t+1}^b), \\ c_t^b &= f_t^b \circ c_{t+1}^b + i_t^b \circ \tanh(W_{xc}^f * \mathcal{H}_t^f + W_{hc}^b * \mathcal{H}_{t+1}^b), \\ \mathcal{H}_t^b &= o_t^b \circ \tanh(c_t^b). \end{aligned} \quad (5.15)$$

Likewise we can formulate the forward ConvLSTM module and the resulting forward-backward features are then combined to get final outputs $\mathcal{Y}_t$ using (5.14).

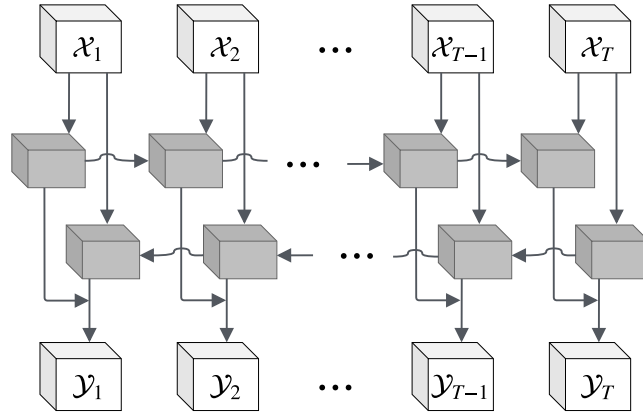


Figure 5.4: Bidirectional operations in the ConvLSTM module

Chapter 6

EXPERIMENTS ON SELECTED APPLICATIONS

Attaining high-performance phoneme recognition is a challenging task when the training data from a degrading channel is limited. Recently, the combination of a convolutional neural network (CNN) and hidden Markov model (HMM) has been used to improve speech recognition performance over the conventional Gaussian mixture model (GMM). Although the CNN/HMM hybrid model is more discriminative, it requires more data to train. In this section, we first address this challenge using a transfer learning in the feature space based on the stacked denoising auto-encoders (SDA). The proposed transfer learning framework populates additional training data for these low-quality sources from abundant and available high-quality sources. In the experimental evaluations, our results within the CNN/HMM hybrid model achieve the phoneme error rates for TM and NB speech as 35.6% and 33.2% respectively.

We then address the adaptation framework to solve food intake classification using teacher-student learning. To learn a better model in a target domain, we use unlabeled parallel data to minimize the loss between distributions. Different from the distillation framework which aims at obtaining small and fast models by imitating the soft output of a larger teacher network, our proposed learning relies on the better representation of target domain when the limitations imposed by the lack of labeled data. We demonstrate that recurrent architecture handle with the classification of intake classes better than baseline CNN models. In other words, recurrent architectures can learn the temporal dynamics depending on their learned parameters. We, therefore, formulate our classification problem as a spatio-temporal sequence learning and use a domain adaptation with the teacher-student framework. Evaluated with the intake classification task, the proposed T/S learning student model can achieve an 83.5% accuracy.

6.1 Phoneme Recognition

6.1.1 Speech Datasets

We utilized the Middle East Technical University Turkish Microphone Speech (METU-TMS) dataset [Salor et al.,] as a large scale close-talk microphone (CM) speech source data $\mathbf{D}_S$. The METU-TMS dataset includes 193 speakers, where each speaker is uttering 40 random TIMIT-like Turkish sentences. Total duration of the METU-TMS is 333 minutes. Recordings are collected with a Sennheiser CM at 16-kHz sampling rate. METU-TMS dataset is phonetically transcribed using the METUbet Turkish phonetic dictionary with 39 phonemes [Salor et al., 2002].

In order to define the low-quality speech sources, we utilized the synchronous TM and CM speech dataset (STCM) that we collected at Koç University [Turan and Erzin, 2016a]. The STCM dataset consists of 800 phonetically-balanced sentences in Turkish, where speech samples are recorded from one male speaker under clean conditions at 16-kHz sampling rate with a total duration of 45 minutes. An IASUS-GP3 headset and Sony condenser tie-pin microphone are used respectively for the TM and CM recordings. Phone level transcripts are extracted through force-alignment using a GMM/HMM system with the METUbet dictionary.

Note that we used the STCM dataset to define the target signal as well as to define the parallel data between the target and source signals. We consider two low-quality speech sources, the throat microphone (TM) and narrow-band (NB) telephony speech recordings. For the TM, target data is defined as the TM recordings of the STCM dataset, $\mathbf{D}_T = \mathbf{D}_T^{TM}$, and the parallel source data is defined as the CM recordings of the STCM dataset, $\mathbf{D}_S^T$.

In order to define the second low-quality speech source, $\mathbf{D}_T = \mathbf{D}_T^{NB}$, the CM recordings of the STCM dataset, $\mathbf{D}_S^T$, is filtered by the modified intermediate reference system (IRS) filter [ITU-T, 2001] to simulate the receiving characteristics of the NB telephony speech. In the NB, the parallel source data is still defined as the CM recordings of the STCM dataset, $\mathbf{D}_S^T$.

6.1.2 Parameters for Feature Extraction and mSDA Calculation

The MFSC features are extracted over 25 ms frames with a stride of 10 ms for $D = 40$ mel-frequency bands. The CNN input, X_c is constructed from the MFSC features over a context window of $B = 15$ frames and then normalized to have zero mean and unit variance. We use 39 class labels corresponding to 39 phonemes in the METUbet Turkish lexicon. A mono-phone language model is estimated from the training set and it is used in the decoding.

In the mSDA feature augmentation, we use AWGN corruption with $R = 9$ repetitions that varies with 5 dB SNR increments from -20 dB to $+20$ dB. The mSDA is constructed with $K = 4$ layers and the regularization parameter is set to $\lambda = 0.01$.

6.1.3 CNN Architecture

Network topology is constructed as 3 convolutional layers interleaved with 2 pooling operations and followed by 2 fully connected (dense) layers. The constructed CNN layers are defined in detail as:

- $\mathcal{L}_1$: 64 filters of (5,5) followed by (2,2) max-pooling and a rectified linear unit (ReLU) activation function
- $\mathcal{L}_2$: 128 filters of (3,3) followed by a ReLU (no pooling)
- $\mathcal{L}_3$: 256 filters (3,3) followed by (2,2) max-pooling and a ReLU
- $\mathcal{L}_4$: 512 hidden units of dense layer followed by a ReLU
- $\mathcal{L}_5$: 39 output units followed by a soft-max activation

For training, the model optimizes cross-entropy loss via mini-batch stochastic gradient descent with a constant learning rate of 0.001. Each batch consists of 32 randomly selected patches from the training data.

6.1.4 Evaluation Metrics

The frame error rate (FER) and phoneme error rate (PER) are used as evaluation metrics. The FER is defined as the percentage of frames that are predicted incorrectly by the CNN classifier. On the other hand, the PER is the phoneme error rate computed at the output of the HMM after the sequence decoding. It is calculated via Levenshtein distance between predicted and ground truth phonemes [Levenshtein, 1966].

6.1.5 Results

We first report PER results of the GMM/HMM system for two target domain data, $\mathbf{D}_T^{TM}$ and $\mathbf{D}_T^{NB}$, in Table 6.1. Note that these systems are trained and tested on the same target data using 5-fold cross-validation. Since the training is performed on the low-quality data, phoneme recognition performs poorly. In general perceived quality of the NB speech is better than the TM speech, and also it has a wider signal bandwidth. This quality difference appears to yield $\sim 4\%$ PER improvement for the NB speech.

The FER and PER results of the CNN/HMM system are presented in Table 6.2. First two experiments in Table 6.2 can be considered as upper-bound performances for the phoneme recognition task, since they run over the high-quality source domain data. Since the METU-TMS dataset is 8-times larger than the STCM, phoneme recognition performance over the $\mathbf{D}_S$ is significantly better than the $\mathbf{D}_S^T$.

Next two experiments in Table 6.2 investigate the baseline performance of the CNN/HMM hybrid system for the two target domain data, $\mathbf{D}_T^{TM}$ and $\mathbf{D}_T^{NB}$. Note

Table 6.1: PER results for the conventional GMM/HMM system

Domain Data	PER
$\mathbf{D}_T^{TM}$	52.1
$\mathbf{D}_T^{NB}$	48.3

Table 6.2: FER and PER results for the CNN/HMM system

Domain Data	<i>FER</i>	<i>PER</i>
$\mathbf{D}_S^T$	24.2	30.4
$\mathbf{D}_S$	20.9	23.6
$\mathbf{D}_T^{TM}$	31.6	40.3
$\mathbf{D}_T^{NB}$	29.7	37.8
$\mathbf{Z}_T^{TM}$	26.2	35.6
$\mathbf{Z}_T^{NB}$	25.4	33.2

that compared to the GMM/HMM system presented in Table 6.1, the hybrid system improves phoneme recognition performance by about relative 10% for both target domains.

Last two experiments in Table 6.2 investigate contribution of the mSDA feature augmentation. Note that in these two experiments, training data includes transformed instances of the source data $\mathbf{Z}_S$, as well as the transformed instances of the target data $\mathbf{Z}_T$ for each low-quality speech sources. However, we again employ 5-fold cross-validation, such that in each fold 80% of the $\mathbf{Z}_T$ is included into the training. The mSDA based transform domain representation $\mathbf{Z}_T$ brings significant improvements, around 5% relative error drop, over the MFSC features $\mathbf{D}_T$ for both low-quality TM and NB speech sources.

6.2 Food Intake Classification

6.2.1 Datasets

The proposed T/S learning for domain adaptation relies on the availability of three datasets, representing the teacher and student domains, and the parallel domain data. The student domain dataset, $\mathcal{D}^S$, is constructed from tracheal TM recordings of 8 subjects, 4 males and 4 females, between 22-29 years old. Intake signals are

collected in laboratory environment with the iASUS NT3 TM at 16 kHz sampling rate. A total of 10 different tasks are recorded including chewing and swallowing of solids (potato chips, cake, biscuits, stick crackers, peanut and chocolate), and swallowing of liquids (water, milk, fizzy drink, and fruit juice). The food types are selected to represent various food textures/categories including wet crispy, dry crispy, hard crunchy and soft, all of which impact the chewing and swallowing sounds. All participants have no history of chewing and swallowing abnormalities. Furthermore, specific intake events are manually annotated within each food cycle by audio-visual inspection of the recorded signals. Each subject visits the laboratory 5 times and consumes these 10 different food items at one recording session. Average duration of each visit for each subject is around 7 min, and the total duration of the student domain dataset $\mathcal{D}^S$ is 276 min.

The teacher domain shall include acoustic CM recordings for the same set of food intake events. However, in the T/S formulation we need a large-size dataset to better train the teacher network. For this purpose, we fetch a large number of ASMR eating event recordings from YouTube to construct the teacher domain dataset $\mathcal{D}^T$. Note that autonomous sensory meridian response, commonly referred as ASMR, is characterized by a combination of positive feelings and a tingling sensation on the skin. There has been growing interest in triggering the ASMR by auditory or visual stimuli that has been created a large collection of ASMR eating videos on YouTube. We collect ASMR eating videos covering the same 10 eating tasks of the student domain dataset $\mathcal{D}^S$, where total recording duration for each task is kept longer than an hour and the total duration of the teacher domain dataset $\mathcal{D}^T$ is formed as 761 min.

As described in section 5.2.3, a parallel domain data, which is recorded synchronously using CM and TM sensors, is needed for the proposed heterogeneous domain adaptation. We collect and use such a parallel CM and TM dataset for TM speech enhancement problem in a recent study [Turan and Erzin,

2016c]. We acquire this dataset as the parallel domain data $\mathcal{D}^P$. Note that the $\mathcal{D}^P$ includes synchronous recordings of 800 phonetically-balanced sentences through a tie-pin CM and IASUS-GP3 TM headset. The total duration of the parallel domain dataset $\mathcal{D}^P$ is 45 min.

6.2.2 Experimental Setup

Food intake events are captured as the acoustic and tissue vibrations through the CM and TM, respectively. Feature representation of these two domain signals are performed by the mel-frequency spectral coefficients (MFSC), which are the log-energies over the mel-scaled critical bands, and they are widely used as the speech signal representations in CNN models [Abdel-Hamid et al., 2012]. The MFSC feature vector f_n at time frame n is extracted over a 25 ms window with a 10 ms frame shift for 40 mel-scaled critical bands.

In this section, we investigate the ConvLSTM and Bi-ConvLSTM networks and compare them with the plain CNN models. The input to the CNN is defined as a $M \times N$ dimensional spectral image of MFSC features over a temporal window centered at time frame n as

$$F_n = [f_{n-\Delta} \cdots f_{n-1}, f_n, f_{n+1}, \cdots f_{n+\Delta}], \quad (6.1)$$

where $M = 40$ is the number of mel-scaled critical bands, $N = 2\Delta + 1$ is the temporal length of the spectral image, and for $\Delta = 7$ the temporal length is chosen as $N = 15$, which corresponds to 150 ms.

The input to the ConvLSTM is defined as a temporal sequence of spectral images as a 3D tensor,

$$T_n = [F_n; F_{n-N}; F_{n-2N}; \cdots; F_{n-(P-1)N}], \quad (6.2)$$

where T_n is $M \times N \times P$ dimensional and includes the sequence of $P = 20$ spectral images ending at time frame n as illustrated in Figure 6.1. Hence, the temporal extent of the T_n corresponds to 3 sec.

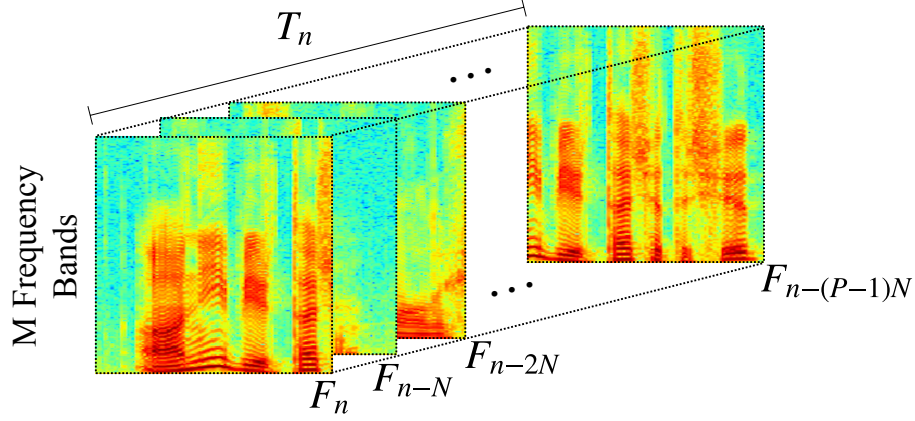


Figure 6.1: Organization of the input speech features to ConvLSTM models

In the proposed ADM system, the CNN, ConvLSTM and Bi-ConvLSTM models are utilized for food intake event classification over the 10 food types. The CNN and ConvLSTM models respectively receives F_n and T_n as inputs with food type labels at time frame n . Furthermore, the models receives K time frames shifted inputs, F_{n+mK} and T_{n+mK} , for $m = 0, 1, 2, \dots$, where we set $K = 30$ in the experimental studies.

Network topology of the CNN is constructed as 3 convolutional layers with 64, 128 and 256 filters of size 5×5 interleaved with 2 pooling operations and followed by 2 fully connected (dense) layers.

We test many variants of the ConvLSTM models with different number of layers. Typically, deeper models can give better results yet for our case the improvement after three layers is not significant. We therefore set the network configuration with three layers of 128, 64, and 64 hidden states respectively. All the input-to-state and state-to-state kernels are of size 5×5 . Networks are optimized by using the stochastic gradient descent algorithm with a mini-batch size of 128. Training process run for 30 epochs with the learning rate initialized to 0.01 with a decay of 0.1 after each 10 epochs.

In the experimental evaluations, we utilized subject independent and subject dependent train/test protocols. In the first protocol, the training and testing is performed using leave-one-subject-out (LOSO) cross-validation to get the subject

independent evaluations. This scheme corresponds to 8-fold cross validation, where in each fold intake events of one subject are tested. In the second protocol, we utilized a 2-fold cross validation. In this setting, we split data into two equal subsets for each subject. Over the stream of classification task, we calculate the true positive, false positive, true negative and false negative rates. Based on these values F-score (F1) and overall accuracy are reported.

6.2.3 Baseline Performances without Domain Adaptation

Baseline accuracy and F-score performances of the CNN, ConvLSTM and Bi-ConvLSTM networks over the teacher and student domain datasets are given in Table 6.3. Note that the T/S learning is not incorporated to the baseline models yet.

We also added the baseline performances of our previously proposed system for model comparison [Turan and Erzin, 2017]. This system is based on a multi-resolution representation of signals using the empirical mode decomposition (EMD) analysis. Intake signals are firstly segmented using the EMD and features are then extracted over the decomposed components. Food intake classification task is carried out using the support vector machines (SVM).

Table 6.3: Baseline performances of the CNN, ConvLSTM and Bi-ConvLSTM topologies (%)

		SVM		CNN		ConvLSTM		Bi-ConvLSTM	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
LOSO	$\mathcal{D}^T$	76.2	73.4	81.8	80.7	87.4	85.1	91.3	88.6
	$\mathcal{D}^S$	65.8	61.2	59.7	58.1	65.4	62.6	67.4	64.9
2-fold	$\mathcal{D}^T$	83.5	80.1	85.4	82.5	89.6	87.3	94.7	92.4
	$\mathcal{D}^S$	71.3	68.6	64.7	61.2	68.4	66.3	70.8	66.5

There is a significant performance difference between the teacher and student domain networks. The performance drop across the domains is mainly due to the weak representational power and scarcity of the student domain TM data. As non-acoustic sensors are not widely used, labelled TM data is scarce and hard to collect. Furthermore, spectral fidelity of the TM sensor is limited compared to the acoustic CM. The TM sensor captures tissue vibration signals, which are typically missing high-frequency components. The low-pass nature of the TM representation is presented with two sample spectrograms in Figure 6.2.

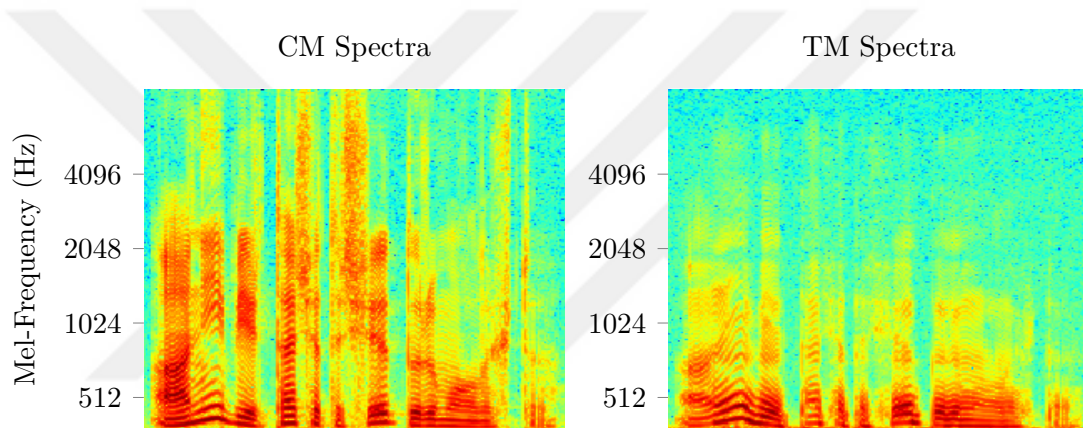


Figure 6.2: Mel-spectrograms of the CM and TM recordings

The baseline performances in Table 6.3 also highlight the benefit of using spatio-temporal modeling in ConvLSTM networks, which outperform the spatial CNN for both the teacher and student domain datasets. Furthermore, bidirectional adaptation, Bi-ConvLSTM, achieves better results than the ConvLSTM. Note that the subject independent LOSO setting has slightly degraded performances, but does not fall off significantly. This is encouraging for the subject independent setting and it's a sign of successful generalization of the food intake classification task.

6.2.4 Performance Analysis with Domain Adaptation

The proposed T/S learning based domain adaptation has been applied using the parallel domain dataset $\mathcal{D}^P$ to improve the representation of the student domain network. Table 6.4 presents the food intake classification performances with the proposed domain adaptation. The accuracy and F-score results have improved above 16% across all settings and networks. This significant improvement ensures functionality of the proposed domain adaptation scheme.

In this new set of experiments, again the 2-fold cross validation yields higher performances than the LOSO protocol, and the Bi-ConvLSTM networks perform consistently better than the others. The best accuracy performances are attained as 83.5% and 87.1% for the subject independent LOSO and the subject dependent 2-fold settings, respectively. Note that the baseline performances with the teacher domain data can be considered as an upper-bound for the intake classification task, since they run over the high-quality and abundant teacher domain data. The best performances of the proposed domain adaptation with the student domain data are reasonably close to these upper-bound baseline results.

Table 6.4: Performances with the proposed T/S Learning based domain adaptation for the student domain $\mathcal{D}^S$ (%)

	CNN		ConvLSTM		Bi-ConvLSTM	
	Acc.	F1	Acc.	F1	Acc.	F1
LOSO	77.1	76.3	81.4	78.6	83.5	82.9
2-fold	81.3	78.9	84.6	82.3	87.1	84.8

Also the best F-score values are attained as 82.9% and 84.8% for the Bi-ConvLSTM network with the subject independent and dependent settings, respectively. Specifically, F-score considers both the precision and the recall of evaluations, as it is calculated from the harmonic average of them. The high value

of F-score indicates how precise the proposed classifier is, as well as how robust it is. Hence, the Bi-ConvLSTM model with the high F-score values is considered as the best performing network for the student domain food intake classification task.

The confusion matrix of the food intake event classification with the Bi-ConvLSTM network is given in Table 6.5 for the LOSO setting. Confusion across solids and liquids is significantly low, on the other hand the liquids seem to attain a similar discrimination as the solids. While the chocolate and cake are the most confusing classes, the best classified solids are potato chips and peanuts, and the liquid is soda.

Table 6.5: Confusion matrix of the Bi-ConvLSTM network with domain adaptation for the LOSO experiment

		Predicted									
		Chips	Cake	Biscuits	Crackers	Peanut	Chocolate	Water	Milk	Soda	Juice
Actual	Chips	92%	1%	3%	2%	2%					
	Cake	4%	73%	5%	3%	6%	8%		1%		
	Biscuits	3%	2%	84%	1%	3%	5%			1%	1%
	Crackers	4%	1%	3%	88%	3%		1%			
	Peanut	2%	1%	3%	3%	90%				1%	
	Chocolate		9%	5%	4%	2%	75%	1%	2%		2%
	Water						1%	85%	3%	6%	5%
	Milk		2%	1%			4%	7 %	79%	4%	3%
	Soda					1%		4%	3%	87%	5%
	Juice		1%	1%			2%	5%	6%	3%	82%

Effect of Adaptation Data

We setup an experiment to investigate the importance of parallel data on the performance of the proposed domain adaptation. For this purpose, we define a combined objective function as

$$\mathcal{L}_{\Gamma, \theta}^c(\mathcal{X}^{T'}, \mathcal{X}^{S'}, \mathcal{Y}) = (1 - \gamma) \cdot \mathcal{L}_{\Gamma, \theta}(\mathcal{X}^{S'}, \mathcal{Y}) + \gamma \cdot \mathcal{L}_{\Gamma, \theta}(\mathcal{X}^{T'}, \mathcal{X}^{S'}, \mathcal{Y}), \quad (6.3)$$

where γ is the combination weight in $[0, 1]$ interval and the first loss function is obtained over the student domain parallel data and defined as

$$\mathcal{L}_{\Gamma, \theta}(\mathcal{X}^{S'}, \mathcal{Y}) = - \sum_{n=1}^N \sum_{j=1}^C p_T(y_n = j | x_n^{S'}; \Gamma) \log p_S(y_n = j | x_n^{S'}; \theta). \quad (6.4)$$

In the experiment, we set the network configuration to Bi-ConvLSTM topology and perform the evaluations with the LOSO setup over the student domain data. Figure 6.3 plots the average LOSO accuracy for the student domain dataset as a function of the weight γ . The performance increases as the weight γ is increasing. Note that $\gamma = 1$ corresponds to the the proposed T/S learning based domain adaptation, where the parallel data is fully utilized. Hence, we can conclude that adding more parallel data samples would improve the performance of the model.

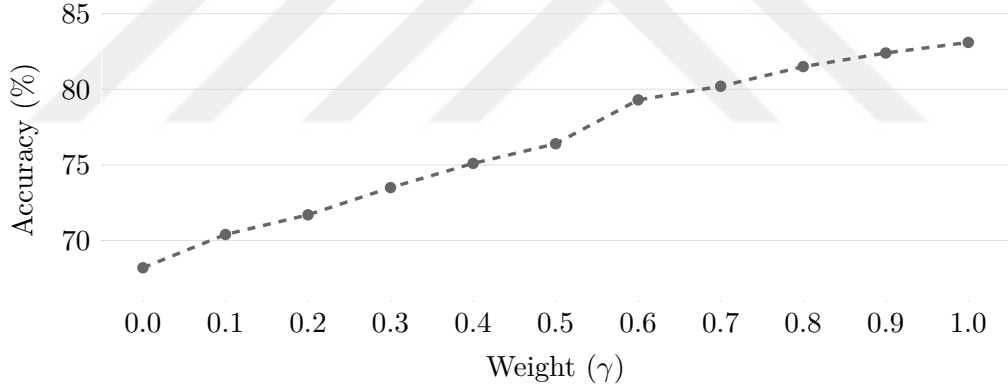


Figure 6.3: Average LOSO accuracy performance as a function of weight γ .

Use of Data Augmentation

In the previous experiment, we observe that incorporation of more parallel samples yields higher classification accuracy within the proposed domain adaptation system. Specifically, the proposed T/S learning can compensate the spectral discrepancy between two modalities with self-supervision of unlabeled parallel data. This enhancement helps the network to capture better representation for the student domain input data. Increasing the available parallel data is expected to

improve generalization capability, as well as the system robustness in the domain adaptation. Hence, we consider to investigate a recently proposed data augmentation technique to grow the parallel domain data.

Typically, data augmentation is performed by either adding noise [Hannun et al., 2014] or by explicit transformation of the data in time domain [Ko et al., 2015] or frequency domain warping [Cui et al., 2015]. In the first case, we need to add an appropriate noise source to the training data. For the latter case, we need an external process that modifies the data prior to the feature extraction.

A data augmentation scheme based on the frequency domain warping, known as vocal tract length perturbation (VTLP), is proposed in [Jaitly and Hinton, 2014]. It demonstrates decent performance improvements in phone error rate for the CNN based speech recognition task on the TIMIT dataset. Vocal track length perturbation can be considered as an appropriate augmentation for the food intake sounds. Hence, we utilized the VTLP frequency domain warping augmentation on the parallel data to improve domain adaptation performance.

We utilized the MFSC features, which are the log-energies over the mel-scaled critical bands. The mel-scale is defined as $m(f) = 1127 \cdot \log(1 + \frac{f}{700})$, where f is the linear frequency variable. The center frequency of the i -th mel-scaled critical band is defined as

$$f(i) = m^{-1} \left(m(F_l) + \frac{m(F_u) - m(F_l)}{M - 1} (i - 1) \right), \quad (6.5)$$

where F_l and F_u are respectively the minimum and maximum frequencies, and $m^{-1}(s)$ is the inverse mel-scale function. Each critical band starts from the center of previous band and ends at the center of next band. We take the minimum and maximum frequencies as $F_l = 0$ Hz and $F_u = 8000$ Hz.

For the VTLP, a random warp factor α is generated for each sample, and the

frequency f is warped to $f' = w(f, \alpha)$ through warping function

$$w(f, \alpha) = \begin{cases} \alpha f & \text{if } f \leq F_h \\ F_u - \frac{F_u - \min(\alpha, 1)F_h}{F_u - \min(\alpha, 1)F_h/\alpha}(F_u - f), & \text{otherwise} \end{cases} \quad (6.6)$$

where F_h is a boundary frequency chosen to cover significant spectral content. We set the boundary frequency $F_h = 4000$ Hz.

In the original VTLP formulation, the warping factor α is taken uniformly for each sample between $[0.8 \ 1.2]$ interval. Therefore, we can double the amount of parallel data using the VTLP. In our experiments, large intervals are observed to create unwanted distortions with higher warping, so we keep the interval smaller and set it to $[0.9 \ 1.1]$.

The warping can be applied directly to the critical bands, where the warped critical band center frequencies are defined as $f'(i) = w(f(i), \alpha)$ for $i = 1, \dots, M$ using (6.6). We end up doubling the parallel domain data in the data augmentation experiments.

Table 6.6: Performances of the proposed domain adaptation using the Bi-ConvLSTM network with the VTLP augmentation (%)

	Before VTLP		After VTLP	
	Accuracy	F1	Accuracy	F1
LOSO	83.5	82.9	85.2	85.4
2-fold	87.1	84.8	89.6	86.3

Table 6.6 presents classification performances before and after the VTLP augmentation with the Bi-ConvLSTM networks. We observe modest performance improvements, 1.7% and 2.5% for the LOSO and 2-fold evaluations, respectively. The VTLP augmentation shows that generating random, linearly warped, variants

of the spectral data can be used to enhance the classification performances. Thus, further gains are achievable using more parallel samples in the proposed T/S learning based domain adaptation.

6.3 Discussion and Conclusion

Wearable sensory solutions for the ADM problem are expected to be robust under different environmental conditions as well as to have practical use in daily life. Our results show that robust and high performing food intake classification is attainable from TM recordings, and hence the TM is a viable sensor and has the capacity to deliver wearable solutions for the ADM.

We formulate the intake classification problem as a spatio-temporal sequence learning and use a new extension of ConvLSTM called Bi-ConvLSTM. The Bi-ConvLSTM network preserves the advantages of the LSTM as well as modelling spatio-temporal data with its inherent convolutional structure. We observed the best intake classification results with the Bi-ConvLSTM network for all experimental settings. This indicates the importance of spatio-temporal modelling for the food intake classification.

We are able to construct a large scale CM dataset from the online ASMR eating event recordings. The CNN and ConvLSTM based networks deliver significantly higher food intake classification accuracy results (above 20% improvements) with the CM dataset compared to the in-house TM dataset. This indicates the richer representational power of the CM recordings, perhaps with better resolution of the high frequency components in the CM recordings. Furthermore, the food intake classification seems to value better spatio-temporal resolution of the CM recordings.

The CM and TM datasets are respectively used to train the teacher and student domain networks. Then, the proposed T/S learning based domain adaptation is successfully utilized to transfer representational capacity of the teacher domain network into the student domain network. Note that the domain adaptation is performed using the parallel CM/TM speech dataset. Classification accuracy improvements are reported above 16% with the domain adaptation. This

indicates the success of the proposed domain adaptation in transferring the better spatio-temporal characterization of the CM recordings to the student network. Furthermore, the speech recordings in the parallel dataset seem to carry information transfer from the CM to TM representations, although they are significantly different from the food intake event recordings.

As demonstrated in the experimental evaluations, increasing the amount of unlabeled parallel data improves the student network better. Hence, we decide to augment the parallel dataset by using the vocal tract length perturbation. After the augmentation, intake classification accuracy reaches 85.2%, which is close to the best teacher domain accuracy at 91.3%. Considering the best classification accuracy of the student domain without domain adaptation, which is 67.4%, the absolute improvement after the domain adaptation with data augmentation is reported as 17.8%.

Future research should address the generalization of current findings for different types of foods, eating contexts, and behaviors. Furthermore, calorie intake estimation should be the long-term target of the ADM systems. Such a target rather emphasizes what and how much a person eats. Although our method is particularly useful at tracking food intake, an accurate estimate of the calorie information based on consumed foods would improve self-control of the subjects as well as the food-related research.

Despite the limitations of this study and opportunities for future research, the current work indicates that the classification of food intake using transfer learning significantly improves the performance of an on-body TM system. Such a food intake monitoring system can allow people to achieve their dietary goals as well as defining new smart and connected health applications.

Chapter 7

CONCLUSIONS

Healthcare applications became widespread after the extensive usage of wearable computation. Dietary monitoring is one of the most important utilization among other healthcare services. Sleep apnea, type II diabetes and other cardiovascular diseases are generally associated with obesity or overweight. Therefore, individuals should track what they eat to minimize the risk of these disorders. Personal records are typically preferred ways of dietary monitoring strategies. However, weight maintenance through self-reports seems highly biased and unreliable because individuals tend to underestimate their food intake. On the other hand, manual logging is regarded as a time-consuming procedure that distracts the users at every meal.

Owing to the recent trend on mobile devices, it is possible to perform reliable and robust monitoring of eating activity through wearable sensors. A novel route in ingestion monitoring research is currently underway and is shaped by the sensing devices. Several sensor systems are investigated for food intake monitoring with single- or multi-sensor configurations using conventional machine learning techniques. Most of the proposed approaches are impractical due to the necessity of multiple sensors specialized for the swallowing or chewing detection separately. Moreover, acoustic modalities are generally vulnerable to environmental noise which results in degradation of performances under daily life conditions. In this thesis, a system based on a single throat microphone sensor positioned on the neck is used for proposed dietary monitoring approaches.

The first step of this thesis investigates the most representative features for the food intake classification task. Recorded signals are firstly segmented and decomposed using the empirical mode decomposition (EMD) analysis. EMD has

been a widely implemented tool to analyze non-stationary and non-linear signals by decomposing data into a series of sub-band oscillations known as intrinsic mode functions (IMFs). For each decomposed IMF signal, time and frequency domain features are computed to provide a multi-resolution representation of the signal. The minimum redundancy maximum relevance principle is then utilized to investigate the most representative features for the food intake classification task using the support vector machines.

An algorithm of automated detection with high accuracy is also needed for a wearable solution. Chewing and swallowing are regarded as the two main stages of food intake mechanism. After the food is taken to the mouth, chewing performs consecutive food degradation to shape a food bolus that can be swallowed. The detection of chewing and swallowing events can be considered in the following part of this thesis. Proposed methodology relies on feature learning in the frequency domain using convolutional neural networks (CNNs). Spectrograms extracted from the recorded food intake sounds are fed into the CNN architecture. This new system does not depend on hand-selected features; instead, it facilitates fine-tuned attributes with neural network structure to achieve superior performance.

Last part of the thesis targets a combined detection and classification of the intake activity events using transfer learning paradigm. Transfer learning has been demonstrated to be effective for many applications as it exploits knowledge present in labeled training data from a source domain to enhance the performance in a target domain, which has little or no labeled data. Heterogeneous transfer learning methods have been developed to address such limitations. This strategy can bridge the feature spaces, data distributions, and other gaps which may be present in these cross-domain learning tasks.

Attaining high-performance phoneme recognition is a challenging task when the training data from a degrading channel is limited. Heterogeneous transfer learning is firstly employed for this task using a transformation in the feature space based on the stacked denoising auto-encoders. The proposed framework populates additional training data for throat microphone recordings from abundant and

available acoustic sources. Secondly, an adaptation framework is proposed to solve the food intake classification using teacher-student learning. To learn a better model in a target domain, this new model uses unlabeled parallel data that minimizes the loss between cross-domain distributions. Different from the original distillation framework, this new learning relies on the better representation of target domain when the limitations imposed by the lack of labeled data.

The main goal of this thesis is to develop non-invasive and automated methods in both detection and classification problems of food intake monitoring. For this purpose, practical and effective solutions via wearable necklace-like sensing are presented here. This research defines the foundations towards the development of an energy efficient, unobtrusive and robust wearable, sensor-based food intake monitoring system. Such a system aims to provide users with quantitative dietary feedback to support improved choices in daily living.

BIBLIOGRAPHY

- [Abdel-Hamid et al., 2014] Abdel-Hamid, O. et al. (2014). Convolutional neural networks for speech recognition. *IEEE Transactions on Audio, Speech & Language Processing*, 22(10):1533–1545.
- [Abdel-Hamid et al., 2012] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., and Penn, G. (2012). Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4277–4280. IEEE.
- [Alam and Bhuiyan, 2013] Alam, S. S. and Bhuiyan, M. I. H. (2013). Detection of seizure and epilepsy using higher order statistics in the emd domain. *IEEE Journal of Biomedical and Health Informatics*, 17(2):312–318.
- [Amft, 2010] Amft, O. (2010). A wearable earpad sensor for chewing monitoring. In *IEEE Sensors Journal*, pages 222–227. IEEE.
- [Amft et al., 2005a] Amft, O., Junker, H., and Troster, G. (2005a). Detection of eating and drinking arm gestures using inertial body-worn sensors. In *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*, pages 160–163. IEEE.
- [Amft et al., 2005b] Amft, O., Stäger, M., Lukowicz, P., and Tröster, G. (2005b). Analysis of chewing sounds for dietary monitoring. In *Ubiquitous Computing*, pages 56–72. Springer.
- [Amft and Troster, 2006] Amft, O. and Troster, G. (2006). Methods for detection and classification of normal swallowing from muscle activation and sound. In *Pervasive Health Conference*, pages 1–10. IEEE.
- [Amft and Tröster, 2008] Amft, O. and Tröster, G. (2008). Recognition of dietary activity events using on-body sensors. *Artificial Intelligence in Medicine*, 42(2):121–136.
- [Arun et al., 2012] Arun, P., Valiyaveetil, M., Biggemann, L., et al. (2012). Modulation of hearing related proteins in the brain and inner ear following repeated blast exposures. *Interventional Medicine and Applied Science*, 4(3):125–131.
- [Aschbrenner et al., 2016] Aschbrenner, K., Naslund, J. A., Shevenell, M., et al. (2016). Feasibility of behavioral weight loss treatment enhanced with peer

- support and mobile health technology for individuals with serious mental illness. *Psychiatric Quarterly*, 87(3):401–415.
- [Aune et al., 2016] Aune, D., Keum, N., Giovannucci, E., Fadnes, L., et al. (2016). Whole grain consumption and risk of cardiovascular disease, cancer, and all cause and cause specific mortality: Systematic review and dose-response meta-analysis of prospective studies. *BMJ*, 353:i2716.
- [Balan et al., 2015] Balan, A. K., Rathod, V., Murphy, K., and Welling, M. (2015). Bayesian dark knowledge. In *Neural Information Processing Systems (NIPS)*, pages 3438–3446.
- [Barthes, 2012] Barthes, R. (2012). Toward a psychosociology of contemporary food consumption. In *Food and Culture*, pages 37–44. Routledge.
- [Belkin et al., 2006] Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(Nov):2399–2434.
- [Bettadapura et al., 2015] Bettadapura, V., Thomaz, E., Parnami, A., Abowd, G., and Essa, I. (2015). Leveraging context to support automated food recognition in restaurants. In *IEEE Winter Conference on Applications of Computer Vision*, pages 580–587. IEEE.
- [Bingham and Cummings, 1983] Bingham, S. and Cummings (1983). The use of 4-aminobenzoic acid as a marker to validate the completeness of 24h urine collections in man. *Clinical Science*, 64(6):629–635.
- [Black and Livingstone, 2003] Black, A. and Livingstone, B. (2003). Markers of the validity of reported energy intake. *The Journal of Nutrition*, 133(3):895S–920S.
- [Blitzer et al., 2006] Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Empirical Methods in Natural Language Processing*, pages 120–128. Association for Computational Linguistics.
- [Borgwardt et al., 2006] Borgwardt, K., Gretton, A., Rasch, M., et al. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57.
- [brown et al., 1998] brown, W. E., Eves, D., Ellison, M., and Braxton, D. (1998). Use of combined electromyography and kinesthesiology during mastication to chart the oral breakdown of foodstuffs. *Journal of Texture Studies*, 29(2):145–167.

- [Brownell and Walsh, 2017] Brownell, K. and Walsh, T. (2017). *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford Publications.
- [Buettner et al., 2002] Buettner, A., Beer, A., Hannig, C., Settles, M., and Schieberle, P. (2002). Physiological and analytical studies on flavor perception dynamics as induced by the eating and swallowing process. *Food Quality and Preference*, 13(7-8):497–504.
- [Burke et al., 2011] Burke, L., Wang, J., and Sevvick, M. A. (2011). Self-monitoring in weight loss: A systematic review of the literature. *Journal of the American Dietetic Association*, 111(1):92–102.
- [Calais Guerra et al., 2011] Calais Guerra, P. H., Veloso, A., Meira Jr, W., and Almeida, V. (2011). From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In *International Conference on Knowledge Discovery and Data Mining*, pages 150–158. ACM.
- [Chan et al., 2015] Chan, W., Ke, N. R., and Lane, I. (2015). Transferring knowledge from a rnn to a dnn. *arXiv preprint arXiv:1504.01483*.
- [Chang et al., 2006] Chang, K.-h., Liu, S.-y., Chu, H.-h., Hsu, J. Y.-j., et al. (2006). The diet-aware dining table: Observing dietary behaviors over a tabletop surface. In *International Conference on Pervasive Computing*, pages 366–382. Springer.
- [Chang et al., 2016] Chang, R. C.-S., Lu, H.-P., Yang, P., and Luarn, P. (2016). Reciprocal reinforcement between wearable activity trackers and social network services in influencing physical activity behaviors. *JMIR mHealth and uHealth*, 4(3).
- [Chebotar and Waters, 2016] Chebotar, Y. and Waters, A. (2016). Distilling knowledge from ensembles of neural networks for speech recognition. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3439–3443.
- [Chen et al., 2012] Chen, M. et al. (2012). Marginalized denoising autoencoders for domain adaptation. In *International Conference on Machine Learning (ICML)*, pages 1627–1634.
- [Chen et al., 2007] Chen, M., Liu, Z., He, L.-W., Chou, P., and Zhang, Z. (2007). Energy-based position estimation of microphones and speakers for ad hoc microphone arrays. In *Applications of Signal Processing to Audio and Acoustics*, pages 22–25. IEEE.
- [Chen et al., 2016] Chen, M., Ma, Y., Song, J., Lai, C.-F., and Hu, B. (2016). Smart clothing: Connecting human with clouds and big data for sustainable health monitoring. *Mobile Networks and Applications*, 21(5):825–845.

- [Chung and Graham, 1997] Chung, F. and Graham, F. C. (1997). *Spectral Graph Theory*. Number 92. American Mathematical Society.
- [Cook et al., 2013] Cook, D., Feuz, K., and Krishnan, N. (2013). Transfer learning for activity recognition: A survey. *Knowledge and Information Systems*, 36(3):537–556.
- [Cordeiro et al., 2015] Cordeiro, F., Epstein, D., Thomaz, E., Bales, E., et al. (2015). Barriers and negative nudges: Exploring challenges in food journaling. In *Annual Conference on Human Factors in Computing Systems*, pages 1159–1162. ACM.
- [Crary et al., 2013] Crary, M. A., Humphrey, J. L., Carnaby-Mann, G., Sambandam, R., Miller, L., and Silliman, S. (2013). Dysphagia, nutrition, and hydration in ischemic stroke patients at admission and discharge from acute care. *Dysphagia*, 28(1):69–76.
- [Cui et al., 2015] Cui, X., Goel, V., and Kingsbury, B. (2015). Data augmentation for deep convolutional neural network acoustic modeling. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4545–4549. IEEE.
- [Dai et al., 2009] Dai, W., Chen, Y., Xue, G.-R., Yang, Q., and Yu, Y. (2009). Translated learning: Transfer learning across different feature spaces. In *Advances in Neural Information Processing Systems (NIPS)*, pages 353–360.
- [Daubechies, 1988] Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996.
- [Daumé III, 2009] Daumé III, H. (2009). Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*.
- [Day and Khoshgoftaar, 2017] Day, O. and Khoshgoftaar, T. M. (2017). A survey on heterogeneous transfer learning. *Journal of Big Data*, 4(1):29.
- [Dekens et al., 2010] Dekens, T., Verhelst, W., Capman, F., and Beaugendre, F. (2010). Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection. In *European Signal Processing Conference (EUSIPCO)*, pages 1978–1982. IEEE.
- [Dhurandhar et al., 2015] Dhurandhar, N., Schoeller, D., Brown, A., Heymsfield, S., et al. (2015). Energy balance measurement: When something is not better than nothing. *International Journal of Obesity*, 39(7):1109.

- [Dietz et al., 2016] Dietz, W., Douglas, C., and Brownson, R. (2016). Chronic disease prevention: Tobacco avoidance, physical activity, and nutrition for a healthy start. *Jama*, 316(16):1645–1646.
- [Ding and Peng, 2005] Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205.
- [Duan et al., 2012] Duan, L., Tsang, I., and Xu, D. (2012). Domain transfer multiple kernel learning. *Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479.
- [Dwyer et al., 2018] Dwyer, J., Coates, P., and Smith, M. (2018). Dietary supplements: Regulatory challenges and research resources. *Nutrients*, 10(1):41.
- [Ertekin et al., 1998] Ertekin, C., Yüceyar, N., et al. (1998). Clinical and electrophysiological evaluation of dysphagia in myasthenia gravis. *Journal of Neurology, Neurosurgery & Psychiatry*, 65(6):848–856.
- [Erzin, 2009a] Erzin, E. (2009a). Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1316–1324.
- [Erzin, 2009b] Erzin, E. (2009b). Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings. *Transactions on Audio, Speech, and Language Processing*, 17(7):1316–1324.
- [Farooq et al., 2014] Farooq, M., Fontana, J., and Sazonov, E. (2014). A novel approach for food intake detection using electroglottography. *Physiological Measurement*, 35(5):739.
- [Fereidooni et al., 2017] Fereidooni, H., Frassetto, T., Miettinen, M., Sadeghi, A.-R., and Conti, M. (2017). Fitness trackers: Fit for health but unfit for security and privacy. In *International Conference on Connected Health*, pages 19–24. IEEE.
- [Fernández-Ballart et al., 2010] Fernández-Ballart, J., Piñol, J. L., Zazpe, I., Corella, D., Carrasco, P., et al. (2010). Relative validity of a semi-quantitative food-frequency questionnaire. *British Journal of Nutrition*, 103(12):1808–1816.
- [Flandrin et al., 2004] Flandrin, P., Rilling, G., and Goncalves, P. (2004). Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 11(2):112–114.
- [Fontana et al., 2015] Fontana, J., Higgins, J., and Schuckers, S. (2015). Energy intake estimation from counts of chews and swallows. *Appetite*, 85:14–21.

- [Forster et al., 2016] Forster, H., Walsh, M., Gibney, M., Brennan, L., and Gibney, E. (2016). Personalised nutrition: The role of new dietary assessment methods. *Proceedings of the Nutrition Society*, 75(1):96–105.
- [Franke et al., 2009] Franke, T., Lukowicz, P., Kunze, K., and Bannach, D. (2009). Can a mobile phone in a pocket reliably recognize ambient sounds? In *International Symposium on Wearable Computers*, pages 161–162. IEEE.
- [Franz et al., 2015] Franz, M., Boucher, J., Rutten-Ramos, S., and VanWormer, J. (2015). Lifestyle weight-loss intervention outcomes in overweight and obese adults. *Journal of the Academy of Nutrition and Dietetics*, 115(9):1447–1463.
- [Freedman et al., 2014] Freedman, L., Commins, J., Moler, J., et al. (2014). Pooled results from validation studies of dietary self-report instruments using recovery biomarkers for energy and protein intake. *American Journal of Epidemiology*, 180(2):172–188.
- [Fukuike et al., 2015] Fukuike, C., Kodama, N., et al. (2015). A novel automated detection system for swallowing sounds during eating and speech under everyday conditions. *Journal of Oral Rehabilitation*, 42(5):340–347.
- [Gemming and Mhurchu, 2016] Gemming, L. and Mhurchu, N. (2016). Dietary under-reporting: What foods and which meals are typically under-reported? *European journal of clinical nutrition*, 70(5):640.
- [Glorot et al., 2011] Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *International Conference on Machine Learning (ICML)*, pages 513–520.
- [Go et al., 2005] Go, V. L., Nguyen, C., Harris, D., and Paul Lee, W.-N. (2005). Nutrient-gene interaction: Metabolic genotype-phenotype relationship. *The Journal of Nutrition*, 135(12):3016S–3020S.
- [Gong et al., 2012] Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 2066–2073. IEEE.
- [Gopalan et al., 2011] Gopalan, R., Li, R., and Chellappa, R. (2011). Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 999–1006. IEEE.
- [Goris and Westerterp, 1999] Goris, A. and Westerterp, K. (1999). Underreporting of habitual food intake is explained by undereating in highly motivated lean women. *The Journal of Nutrition*, 129(4):878–882.

- [Goris and Westerterp, 2000] Goris, A. H. and Westerterp, K. R. (2000). Improved reporting of habitual food intake after confrontation with earlier results on food reporting. *British Journal of Nutrition*, 83(4):363–369.
- [Graves et al., 2013] Graves, A., Jaitly, N., and Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 273–278. IEEE.
- [Greff et al., 2017] Greff, K., Srivastava, R., et al. (2017). Lstm: A search space odyssey. *Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232.
- [Gu et al., 2005] Gu, D., Reynolds, K., Wu, X., Chen, J., Duan, X., et al. (2005). Prevalence of the metabolic syndrome and overweight among adults in china. *The Lancet*, 365(9468):1398–1405.
- [Gudupudi et al., 2015] Gudupudi, L. K., Chatlani, N., Beaugeant, C., and Evans, N. (2015). Non-linear acoustic echo cancellation using empirical mode decomposition. In *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pages 614–618. IEEE.
- [Hammerla et al., 2016] Hammerla, N., Halloran, S., and Ploetz, T. (2016). Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*.
- [Hannun et al., 2014] Hannun, A., Case, C., Casper, J., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.
- [Harari et al., 2001] Harari, D., Furst, M., Kiryati, N., Caspi, A., and Davidson, M. (2001). A computer-based method for the assessment of body-image distortions in anorexia-nervosa patients. *IEEE Transactions on Information Technology in Biomedicine*, 5(4):311–319.
- [Harel and Mannor, 2010] Harel, M. and Mannor, S. (2010). Learning from multiple outlooks. *arXiv preprint arXiv:1005.0027*.
- [Hatori et al., 2012] Hatori, M., Vollmers, C., Zarrinpar, A., et al. (2012). Time-restricted feeding without reducing caloric intake prevents metabolic diseases in mice fed a high-fat diet. *Cell Metabolism*, 15(6):848–860.
- [Heath, 2002] Heath, R. (2002). The oral management of food: The bases of oral success and for understanding the sensations that drive us to eat. *Food Quality and Preference*, 13(7-8):453–461.

- [Hinton et al., 2012a] Hinton, G., Deng, L., et al. (2012a). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.
- [Hinton et al., 2012b] Hinton, G., Srivastava, N., et al. (2012b). Improving neural networks by preventing co-adaptation of feature detectors. *Arxiv Preprint 1207.0580*.
- [Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- [Huang et al., 2007] Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2007). Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 601–608.
- [Huang et al., 1998] Huang, N. E., Shen, Z., et al. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 454, pages 903–995. The Royal Society.
- [Huse et al., 1974] Huse, D., Nelson, R., Briones, E., and Hodgson, P. (1974). Urinary nitrogen excretion as objective measure of dietary intake. *The American Journal of Clinical Nutrition*, 27(8):771–773.
- [IDC, 2018] IDC (2018). Worldwide wearables market ticks up 5.5% due to gains in emerging markets. <https://www.idc.com/getdoc.jsp?containerId=prUS44247418>. Accessed: 2018-12-20.
- [ITU-T, 2001] ITU-T (2001). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Recommendation P.862*.
- [Jacobs, 2012] Jacobs, D. R. (2012). Challenges in research in nutritional epidemiology. In *Nutritional Health*, pages 29–42. Springer.
- [Jaitly and Hinton, 2014] Jaitly, N. and Hinton, G. (2014). Vocal tract length perturbation improves speech recognition. In *International Conference on Machine Learning (ICML)*, volume 117.
- [Jakubowicz et al., 2013] Jakubowicz, D., Barnea, M., Wainstein, J., and Froy, O. (2013). High caloric intake at breakfast vs. dinner differentially influences weight loss of overweight and obese women. *Obesity*, 21(12):2504–2512.
- [Jéquier and Tappy, 1999] Jéquier, E. and Tappy, L. (1999). Regulation of body weight in humans. *Physiological Reviews*, 79(2):451–480.

- [Jiang and Zhai, 2007] Jiang, J. and Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *Annual Meeting of the Association of Computational Linguistics*, pages 264–271.
- [Kadomura et al., 2013] Kadomura, A., Li, C.-Y., Chen, Y.-C., Chu, H.-H., et al. (2013). Sensing fork and persuasive game for improving eating behavior. In *Pervasive and Ubiquitous Computing*, pages 71–74. ACM.
- [Kawano and Yanai, 2014] Kawano, Y. and Yanai, K. (2014). Food image recognition with deep convolutional features. In *Pervasive and Ubiquitous Computing*, pages 589–593. ACM.
- [Kearney, 2010] Kearney, J. (2010). Food consumption trends and drivers. *Philosophical Transactions of the Royal Society: Biological Sciences*, 365(1554):2793–2807.
- [Kim et al., 2012] Kim, H.-J., Kim, M., Lee, S.-J., and Choi, Y. S. (2012). An analysis of eating activities for automatic food type recognition. In *Signal & Information Processing Association Annual Conference (APSIPA)*, pages 1–5. IEEE.
- [Kina-Tanada et al., 2017] Kina-Tanada, M., Sakanashi, M., Tanimoto, A., et al. (2017). Long-term dietary nitrite and nitrate deficiency causes the metabolic syndrome, endothelial dysfunction and cardiovascular death in mice. *Diabetologia*, 60(6):1138–1151.
- [Klap and Shinar, 2013] Klap, T. and Shinar, Z. (2013). Using piezoelectric sensor for continuous-contact-free monitoring of heart and respiration rates in real-life hospital settings. In *Computing in Cardiology Conference (CinC), 2013*, pages 671–674. IEEE.
- [Ko et al., 2015] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA.
- [Kohyama et al., 2003] Kohyama, K., Mioche, L., and Bourdio3, P. (2003). Influence of age and dental status on chewing behaviour studied by emg recordings during consumption of various food samples. *Gerodontology*, 20(1):15–23.
- [Kohyama et al., 2007] Kohyama, K., Nakayama, Y., Yamaguchi, I., Yamaguchi, M., Hayakawa, F., and Sasaki, T. (2007). Mastication efforts on block and finely cut foods studied by electromyography. *Food Quality and Preference*, 18(2):313–320.

- [Kojima et al., 2016] Kojima, T., Ijiri, T., White, J., Kataoka, H., and Hirabayashi, A. (2016). Cogknife: Food recognition from their cutting sounds. In *International Conference on Multimedia & Expo (ICME)*, pages 1–6. IEEE.
- [Konttinen et al., 2010] Konttinen, H., Männistö, S., Sarlio-Lähteenkorva, S., Silventoinen, K., and Haukkala, A. (2010). Emotional eating, depressive symptoms and self-reported food consumption. a population-based study. *Appetite*, 54(3):473–479.
- [Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105.
- [Kulis et al., 2011] Kulis, B., Saenko, K., and Darrell, T. (2011). What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1785–1792. IEEE.
- [Kurth and Schaffrath, 2007] Kurth, B. and Schaffrath, R. (2007). The prevalence of overweight and obese children and adolescents living in germany. *Bundesgesundheitsblatt*, 50(5-6):736–743.
- [Labonté et al., 2012] Labonté, M.-È., Cyr, A., Baril-Gravel, L., Royer, M., and Lamarche, B. (2012). Validity and reproducibility of a web-based, self-administered food frequency questionnaire. *European journal of clinical nutrition*, 66(2):166–173.
- [Lansky and Brownell, 1982] Lansky, D. and Brownell, K. D. (1982). Estimates of food quantity and calories: errors in self-report among obese patients. *The American journal of clinical nutrition*, 35(4):727–732.
- [Lara et al., 2013] Lara, O., Labrador, M., et al. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3):1192–1209.
- [Larson et al., 2011] Larson, E., Lee, T., et al. (2011). Accurate and privacy preserving cough sensing using a low-cost microphone. In *ACM Conference on Ubiquitous Computing*, pages 375–384.
- [Lehnert et al., 2013] Lehnert, T., Sonntag, D., Konnopka, A., Riedel-Heller, S., and König, H.-H. (2013). Economic costs of overweight and obesity. *Best Practice & Research Clinical Endocrinology & Metabolism*, 27(2):105–115.
- [Levenshtein, 1966] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Physics*, 10(8):707–710.

- [Li et al., 2017] Li, J., Seltzer, M., Wang, X., Zhao, R., and Gong, Y. (2017). Large-scale domain adaptation via teacher-student learning. *arXiv preprint arXiv:1708.05466*.
- [Liu et al., 2012] Liu, J., Johns, E., Atallah, L., Pettitt, C., et al. (2012). An intelligent food-intake monitoring system using wearable sensors. In *International Conference on Wearable and Implantable Body Sensor Networks*, pages 154–160. IEEE.
- [Livingstone et al., 1990] Livingstone, M., Prentice, A., Strain, J., Coward, W., Black, A., Barker, M., McKenna, P., and Whitehead, R. (1990). Accuracy of weighed dietary records in studies of diet and health. *Bmj*, 300(6726):708–712.
- [Livingstone and Black, 2003] Livingstone, M. B. E. and Black, A. E. (2003). Markers of the validity of reported energy intake. *The Journal of Nutrition*, 133(3):895S–920S.
- [Long et al., 2014] Long, M., Wang, J., Ding, G., Pan, S. J., and Philip, Y. (2014). Adaptation regularization: A general framework for transfer learning. *Transactions on Knowledge and Data Engineering*, 26(5):1076–1089.
- [Long et al., 2013] Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. (2013). Transfer feature learning with joint distribution adaptation. In *International Conference on Computer Vision*, pages 2200–2207.
- [Lu et al., 2009] Lu, H., Pan, W., Lane, N., Choudhury, T., and Campbell, A. (2009). Soundsense: Scalable sound sensing for people-centric applications on mobile phones. In *international Conference on Mobile Systems, Applications, and Services*, pages 165–178. ACM.
- [Ma et al., 2015] Ma, X., Tao, Z., Wang, Y., Yu, H., and Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research: Emerging Technologies*, 54:187–197.
- [Ma et al., 2012] Ma, Y., Luo, G., Zeng, X., and Chen, A. (2012). Transfer learning for cross-company software defect prediction. *Information and Software Technology*, 54(3):248–256.
- [Maas et al., 2013] Maas, A., Hannun, A., and Ng, A. (2013). Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML)*.
- [Maekawa, 2013] Maekawa, T. (2013). A sensor device for automatic food lifelogging that is embedded in home ceiling light: A preliminary investigation. In

- International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 405–407. IEEE.
- [Makeyev et al., 2012] Makeyev, O., Lopez-Meyer, P., et al. (2012). Automatic food intake detection based on swallowing sounds. *Biomedical Signal Processing and Control*, 7(6):649–656.
- [Malik et al., 2013] Malik, V., Willett, W., and Hu, F. (2013). Global obesity: Trends, risk factors and policy implications. *Nature Reviews Endocrinology*, 9(1):13.
- [Mandic et al., 2013] Mandic, D. P., ur Rehman, N., Wu, Z., and Huang, N. E. (2013). Empirical mode decomposition-based time-frequency analysis of multivariate signals: The power of adaptive data analysis. *IEEE signal processing magazine*, 30(6):74–86.
- [Mao et al., 2014] Mao, Q., Dong, M., et al. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213.
- [Marder et al., 2004] Marder, S., Essock, S., Miller, A., Buchanan, R., et al. (2004). Physical health monitoring of patients with schizophrenia. *American Journal of Psychiatry*, 161(8):1334–1349.
- [Marshall et al., 2001] Marshall, T., Stumbo, P., Warren, J., and Xie, X.-J. (2001). Inadequate nutrient intakes are common and are associated with low diet variety in rural, community-dwelling elderly. *The Journal of Nutrition*, 131(8):2192–2196.
- [Martin, 2005] Martin, R. (2005). Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Transactions on Audio, Speech & Language Processing*, 13(5):845–856.
- [Mohamed et al., 2012] Mohamed, A.-R., Dahl, G. E., Hinton, G., et al. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech & Language Processing*, 20(1):14–22.
- [Murakami et al., 2016] Murakami, H., Kawakami, R., Nakae, S., Nakata, Y., et al. (2016). Accuracy of wearable devices for estimating total energy expenditure. *JAMA internal medicine*, 176(5):702–703.
- [Ng et al., 2014] Ng, M., Fleming, T., Robinson, M., et al. (2014). Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013. *The Lancet*, 384(9945):766–781.

- [Nguyen et al., 2014] Nguyen, X. V., Chan, J., Romano, S., and Bailey, J. (2014). Effective global approaches for mutual information based feature selection. In *Proceedings of the international conference on knowledge discovery and data mining*, pages 512–521. ACM.
- [Nishimura and Kuroda, 2008] Nishimura, J. and Kuroda, T. (2008). Eating habits monitoring using wireless wearable in-ear microphone. In *International Symposium on Wireless Pervasive Computing*, pages 130–132. IEEE.
- [Olubanjo and Ghovanloo, 2014] Olubanjo, T. and Ghovanloo, M. (2014). Real-time swallowing detection based on tracheal acoustics. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4384–4388. IEEE.
- [Palaz et al., 2015] Palaz, D., Magimai-Doss, M., and Collobert, R. (2015). Convolutional neural networks-based continuous speech recognition using raw speech signal. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4295–4299. IEEE.
- [Pan et al., 2008] Pan, S. J., Kwok, J., and Yang, Q. (2008). Transfer learning via dimensionality reduction. In *AAAI Conference on Artificial Intelligence*, volume 8, pages 677–682.
- [Pan et al., 2010] Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., and Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *International Conference on World Wide Web*, pages 751–760. ACM.
- [Pan et al., 2011] Pan, S. J., Tsang, I., Kwok, J., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *Transactions on Neural Networks*, 22(2):199–210.
- [Papapanagiotou et al., 2017] Papapanagiotou, V., Diou, C., and Delopoulos, A. (2017). Chewing detection from an in-ear microphone using convolutional neural networks. In *International Conference of the Engineering in Medicine and Biology Society*, pages 1258–1261. IEEE.
- [Pascanu et al., 2013] Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pages 1310–1318.
- [Pascual et al., 2018] Pascual, S., Park, M., Serra, J., Bonafonte, A., and Ahn, K. (2018). Language and noise transfer in speech enhancement generative adversarial network. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5019–5023.

- [Päßler and Fischer, 2014] Päßler, S. and Fischer, W.-J. (2014). Food intake monitoring: Automated chew event detection in chewing sounds. *Journal of Biomedical and Health Informatics*, 18(1):278–289.
- [Päßler et al., 2012] Päßler, S., Wolff, M., and Fischer, W.-J. (2012). Food intake monitoring: An acoustical approach to automated food intake activity detection and classification of consumed food. *Physiological Measurement*, 33(6):1073.
- [Patel et al., 2017] Patel, M., Foschini, L., Kurtzman, G., et al. (2017). Using wearable devices and smartphones to track physical activity. *Annals of Internal Medicine*, 167(10):755–757.
- [Poppitt et al., 1998] Poppitt, S., Swann, D., Black, A., and Prentice, A. (1998). Assessment of selective under-reporting of food intake by both obese and non-obese women in a metabolic facility. *International Journal of Obesity*, 22(4):303.
- [Pozza and Isidori, 2018] Pozza, C. and Isidori, A. (2018). What’s behind the obesity epidemic. In *Imaging in Bariatric Surgery*, pages 1–8. Springer.
- [Prettenhofer and Stein, 2010] Prettenhofer, P. and Stein, B. (2010). Cross-language text classification using structural correspondence learning. In *Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127.
- [Qi et al., 2011] Qi, G.-J., Aggarwal, C., and Huang, T. (2011). Towards semantic knowledge propagation from text corpus to web images. In *International Conference on World Wide Web*, pages 297–306. ACM.
- [Rabbi et al., 2015] Rabbi, M., Pfammatter, A., Zhang, M., Spring, B., and Choudhury, T. (2015). Automated personalized feedback for physical activity and dietary behavior change with mobile phones: A randomized controlled trial on adults. *JMIR mHealth and uHealth*, 3(2).
- [Raina et al., 2007] Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. (2007). Self-taught learning: Transfer learning from unlabeled data. In *International Conference on Machine Learning (ICML)*, pages 759–766. ACM.
- [Rateni et al., 2017] Rateni, G., Dario, P., and Cavallo, F. (2017). Smartphone-based food diagnostic technologies: A review. *Sensors*, 17(6):1453.
- [Romero et al., 2014] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2014). Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

- [Rossi et al., 2013] Rossi, M., Feese, S., Amft, O., Braune, N., Martis, S., and Tröster, G. (2013). Ambientsense: A real-time ambient sound recognition system for smartphones. In *International Conference on Pervasive Computing and Communications*, pages 230–235. IEEE.
- [Rusin et al., 2013] Rusin, M., Årsand, E., and Hartvigsen, G. (2013). Functionalities and input methods for recording food intake: A systematic review. *International Journal of Medical Informatics*, 82(8):653–664.
- [Sainath et al., 2012] Sainath, T. N. et al. (2012). Auto-encoder bottleneck features using deep belief networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4153–4156. IEEE.
- [Salor et al.,] Salor, Ö. et al. Middle East Technical University Turkish Microphone Speech. LDC2006S33. Philadelphia: Linguistic Data Consortium, 2006.
- [Salor et al., 2002] Salor, Ö. et al. (2002). On developing new text and audio corpora and speech recognition tools for the Turkish language. In *International Conference on Spoken Language Processing*.
- [Satia-Abouta et al., 2002] Satia-Abouta, J., Patterson, R., Neuhouser, M., and Elder, J. (2002). Dietary acculturation: Applications to nutrition research and dietetics. *Journal of the American Dietetic Association*, 102(8):1105–1118.
- [Sato and Nakashima, 2006] Sato, K. and Nakashima, T. (2006). Human adult deglutition during sleep. *Annals of Otology, Rhinology & Laryngology*, 115(5):334–339.
- [Sazonov et al., 2010] Sazonov, E., Makeyev, O., et al. (2010). Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior. *Transactions on Biomedical Engineering*, 57(3):626–633.
- [Sazonov and Fontana, 2012] Sazonov, E. S. and Fontana, J. M. (2012). A sensor system for automatic detection of food intake through non-invasive monitoring of chewing. *IEEE sensors journal*, 12(5):1340–1348.
- [Schoeller et al., 1990] Schoeller, D., Bandini, L., and Dietz, W. (1990). Inaccuracies in self-reported intake identified by comparison with the doubly-labelled water method. *Canadian Journal of Physiology and Pharmacology*, 68(7):941–949.
- [Schwartz et al., 2000] Schwartz, M. W., Woods, S. C., Porte Jr, D., Seeley, R. J., and Baskin, D. G. (2000). Central nervous system control of food intake. *Nature*, 404(6778):661.

- [Sharp and Allman-Farinelli, 2014] Sharp, D. and Allman-Farinelli, M. (2014). Feasibility and validity of mobile phones to assess dietary intake. *Nutrition*, 30(11-12):1257–1266.
- [Sharp and Sobal, 2012] Sharp, D. and Sobal, J. (2012). Using plate mapping to examine sensitivity to plate size in food portions and meal composition among college students. *Appetite*, 59(3):639–645.
- [Shi et al., 2010] Shi, X., Liu, Q., Fan, W., Philip, Y., and Zhu, R. (2010). Transfer learning on heterogenous feature spaces via spectral transformation. In *international Conference on Data Mining*, pages 1049–1054. IEEE.
- [Shi and Sha, 2012] Shi, Y. and Sha, F. (2012). Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. *arXiv preprint arXiv:1206.6438*.
- [Shuzo et al., 2010] Shuzo, M., Komori, S., Takashima, T., et al. (2010). Wearable eating habit sensing system using internal body sound. *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, 4(1):158–166.
- [Simmons and Reuben, 2000] Simmons, S. and Reuben, D. (2000). Nutritional intake monitoring for nursing home residents: A comparison of staff documentation, direct observation, and photography methods. *Journal of the American Geriatrics Society*, 48(2):209–213.
- [Singer et al., 2014] Singer, P., Hiesmayr, M., Biolo, G., Felbinger, T., et al. (2014). Pragmatic approach to nutrition in the icu: Expert opinion regarding which calorie protein target. *Clinical Nutrition*, 33(2):246–251.
- [Skinner et al., 2016] Skinner, A. C., Perrin, E., and Skelton, J. (2016). Prevalence of obesity and severe obesity in us children, 1999-2014. *Obesity*, 24(5):1116–1123.
- [Smith and Smith, 2016] Smith, K. B. and Smith, M. S. (2016). Obesity statistics. *Primary Care: Clinics in Office Practice*, 43(1):121–135.
- [Soleymani et al., 2016] Soleymani, T., Daniel, S., and Garvey, T. (2016). Weight maintenance: Challenges, tools and strategies for primary care physicians. *Obesity Reviews*, 17(1):81–93.
- [Srinivasan et al., 2017] Srinivasan, B., Lee, S., Erickson, D., and Mehta, S. (2017). Precision nutrition—review of methods for point-of-care assessment of nutritional status. *Current Opinion in Biotechnology*, 44:103–108.
- [Stein and Brooks, 2017] Stein, N. and Brooks, K. (2017). A fully automated conversational artificial intelligence for weight loss: Longitudinal observational study among overweight and obese adults. *JMIR Diabetes*, 2(2):e28.

- [Steinwart, 2001] Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93.
- [Stellar and Shrager, 1985] Stellar, E. and Shrager, E. (1985). Chews and swallows and the microstructure of eating. *The American Journal of Clinical Nutrition*, 42(5):973–982.
- [Sun and Saenko, 2016] Sun, B. and Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer.
- [Sun et al., 2014] Sun, M., Burke, L., Mao, Z.-H., Chen, Y., Chen, H.-C., et al. (2014). ebutton: A wearable computer for health monitoring and personal assistance. In *Design Automation Conference*, pages 1–6. ACM.
- [Sundaravadivel et al., 2018] Sundaravadivel, P., Kesavan, K., Kesavan, L., Mohanty, S., and Kougianos, E. (2018). Smart-log: A deep-learning based automated nutrition monitoring system in the iot. *IEEE Transactions on Consumer Electronics*, 64(3):390–398.
- [Sung et al., 2005] Sung, M., Marci, C., and Pentland, A. (2005). Wearable feedback systems for rehabilitation. *Journal of Neuroengineering and Rehabilitation*, 2(1):17.
- [Tanaka et al., 2012] Tanaka, N., Nohara, K., Okuno, et al. (2012). Development of a swallowing frequency meter using a laryngeal microphone. *Journal of Oral Rehabilitation*, 39(6):411–420.
- [Tang et al., 2018] Tang, Y., Zou, W., Jin, Z., and Li, X. (2018). Multi-scale spatiotemporal conv-lstm network for video saliency detection. In *International Conference on Multimedia Retrieval*, pages 362–369. ACM.
- [Tanigawa et al., 2008] Tanigawa, S., Nishihara, H., Kaneda, S., and Haga, H. (2008). Detecting mastication by using microwave doppler sensor. In *international Conference on Pervasive Technologies Related to Assistive Environments*, page 88. ACM.
- [Tanumihardjo et al., 2007] Tanumihardjo, S., Anderson, C., Kaufer-Horwitz, M., et al. (2007). Poverty, obesity, and malnutrition: An international perspective recognizing the paradox. *Journal of the American Dietetic Association*, 107(11):1966–1972.
- [Thomaz et al., 2013] Thomaz, E., Parnami, A., Essa, I., and Abowd, G. (2013). Feasibility of identifying eating moments from first-person images

- leveraging human computation. In *International SenseCam & Pervasive Imaging Conference*, pages 26–33. ACM.
- [Thompson and Subar, 2017] Thompson, F. and Subar, A. (2017). Dietary assessment methodology. In *Nutrition in the Prevention and Treatment of Disease*, pages 5–48. Elsevier.
- [Troiano, 2006] Troiano, R. (2006). Translating accelerometer counts into energy expenditure: Advancing the quest. *Journal of Applied Physiology*, 100(4):1107–1108.
- [Tür et al., 2012] Tür, G., Deng, L., Hakkani-Tür, D., and He, X. (2012). Towards deeper understanding: Deep convex networks for semantic utterance classification. In *Acoustics, Speech and Signal Processing Conference (ICASSP)*, pages 5045–5048.
- [Turan, 2013] Turan, M. A. T. (2013). Enhancement of throat microphone recordings using gaussian mixture model probabilistic estimator. *arXiv preprint arXiv:1804.05937*.
- [Turan and Erzin, 2013] Turan, M. A. T. and Erzin, E. (2013). Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra. In *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pages 7049–7053. IEEE.
- [Turan and Erzin, 2016a] Turan, M. A. T. and Erzin, E. (2016a). Food intake classification using throat microphone. In *Signal Processing and Communications Applications Conference*, pages 1873–1876. IEEE.
- [Turan and Erzin, 2016b] Turan, M. A. T. and Erzin, E. (2016b). Source and filter estimation for throat-microphone speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):265–275.
- [Turan and Erzin, 2016c] Turan, M. A. T. and Erzin, E. (2016c). Source and filter estimation for throat-microphone speech enhancement. *Transactions on Audio, Speech, and Language Processing*, 24(2):265–275.
- [Turan and Erzin, 2017] Turan, M. A. T. and Erzin, E. (2017). Empirical mode decomposition of throat microphone recordings for intake classification. In *International Workshop on Multimedia for Personal Health and Health Care*, pages 45–52. ACM.
- [Turan and Erzin, 2018] Turan, M. A. T. and Erzin, E. (2018). Detection of food intake events from throat microphone recordings using convolutional neural networks. In *International Conference on Multimedia & Expo*, pages 1–6. IEEE.

- [Tylka et al., 2014] Tylka, T., Annunziato, R., Burgard, D., Daníelsdóttir, S., et al. (2014). The weight-inclusive versus weight-normative approach to health: Evaluating the evidence for prioritizing well-being over weight loss. *Journal of Obesity*, 2014.
- [Van Gaal et al., 2006] Van Gaal, L. F., Mertens, I. L., and Christophe, E. (2006). Mechanisms linking obesity with cardiovascular disease. *Nature*, 444(7121):875–880.
- [Vermeir and Verbeke, 2006] Vermeir, I. and Verbeke, W. (2006). Sustainable food consumption: Exploring the consumer “attitude–behavioral intention” gap. *Journal of Agricultural and Environmental Ethics*, 19(2):169–194.
- [Vilar-Gomez et al., 2015] Vilar-Gomez, E., Martinez-Perez, Y., Calzadilla-Bertot, L., et al. (2015). Weight loss through lifestyle modification significantly reduces features of nonalcoholic steatohepatitis. *Gastroenterology*, 149(2):367–378.
- [Visscher and Seidell, 2001] Visscher, T. and Seidell, J. (2001). The public health impact of obesity. *Annual Review of Public Health*, 22(1):355–375.
- [Vu et al., 2017] Vu, T., Lin, F., Alshurafa, N., and Xu, W. (2017). Wearable food intake monitoring technologies: A comprehensive review. *Computers*, 6(1):4.
- [Walker and Bhatia, 2014] Walker, W. and Bhatia, D. (2014). Automated ingestion detection for a health monitoring system. *Journal of Biomedical and Health Informatics*, 18(2):682–692.
- [Wang and Mahadevan, 2011] Wang, C. and Mahadevan, S. (2011). Heterogeneous domain adaptation using manifold alignment. In *International Joint Conference on Artificial Intelligence*.
- [Wang et al., 2006] Wang, D.-H., Kogashiwa, M., and Kira, S. (2006). Development of a new instrument for evaluating individuals’ dietary intakes. *Journal of the American Dietetic Association*, 106(10):1588–1593.
- [Wang et al., 2018] Wang, W., Shen, J., and Shao, L. (2018). Video salient object detection via fully convolutional networks. *Transactions on Image Processing*, 27(1):38–49.
- [Wei and Pal, 2011] Wei, B. and Pal, C. (2011). Heterogeneous transfer learning with rbms. In *AAAI Conference on Artificial Intelligence*.
- [Weiss et al., 2016] Weiss, K., Khoshgoftaar, T., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1):1–40.

- [Weninger et al., 2015] Weninger, F., Erdogan, H., et al. (2015). Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 91–99. Springer.
- [Westerterp et al., 1986] Westerterp, S., Es, V., and Hoor, T. (1986). Use of the doubly labeled water technique in humans during heavy sustained exercise. *Journal of Applied Physiology*, 61(6):2162–2167.
- [Wharton et al., 2014] Wharton, C., Johnston, C., Cunningham, B., and Sterner, D. (2014). Dietary self-monitoring improves with use of smartphone app technology. *Journal of Nutrition Education and Behavior*, 46(5):440–444.
- [Wilde and Garvin, 2007] Wilde, M. and Garvin, S. (2007). A concept analysis of self-monitoring. *Journal of Advanced Nursing*, 57(3):339–350.
- [Woda et al., 2006] Woda, A., Foster, K., Mishellany, A., and Peyron, M. (2006). Adaptation of healthy mastication to factors pertaining to the individual or to the food. *Physiology & Behavior*, 89(1):28–35.
- [Wu et al., 2011] Wu, X., Xu, D., Duan, L., and Luo, J. (2011). Action recognition using context and appearance distribution features. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 489–496. IEEE.
- [Xingjian et al., 2015] Xingjian, S., Chen, Z., Wang, H., et al. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NIPS)*, pages 802–810.
- [Xiong et al., 2018] Xiong, W., Wu, L., Allewa, F., Droppo, J., Huang, X., and Stolcke, A. (2018). The Microsoft 2017 conversational speech recognition system. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5934–5938. IEEE.
- [Xu et al., 2014] Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2014). Cross-language transfer learning for deep neural network based speech enhancement. In *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 336–340. IEEE.
- [Yang and Colditz, 2015] Yang, L. and Colditz, G. (2015). Prevalence of overweight and obesity in the united states. *JAMA Internal Medicine*, 175(8):1412–1413.
- [Yatani and Truong, 2012] Yatani, K. and Truong, K. N. (2012). Bodyscope: A wearable acoustic sensor for activity recognition. In *International Conference on Ubiquitous Computing*, pages 341–350. ACM.

- [Yi et al., 2018] Yi, J., Tao, J., Wen, Z., and Liu, B. (2018). Distilling knowledge using parallel data for far-field speech recognition. *arXiv preprint arXiv:1802.06941*.
- [Yim et al., 2017] Yim, J., Joo, D., Bae, J., and Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4133–4141.
- [Yon et al., 2007] Yon, B., Johnson, R. K., Harvey-Berino, J., Gold, B. C., and Howard, A. (2007). Personal digital assistants are comparable to traditional diaries for dietary self-monitoring during a weight loss program. *Journal of Behavioral Medicine*, 30(2):165–175.
- [Yu et al., 2013] Yu, D., Seltzer, M., Li, J., et al. (2013). Feature learning in deep neural networks-studies on speech recognition tasks. *Arxiv Preprint 1301.3605*.
- [Zen and Sak, 2015] Zen, H. and Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Acoustics, Speech and Signal Processing Conference (ICASSP)*.
- [Zhang et al., 2009] Zhang, S., Ang, M. H., Xiao, W., and Tham, C. K. (2009). Detection of activities by wireless sensors for daily life surveillance: Eating and drinking. *Sensors*, 9(3):1499–1517.
- [Zhou et al., 2014] Zhou, J. T. et al. (2014). Hybrid heterogeneous transfer learning through deep learning. In *AAAI Conference on Artificial Intelligence*.
- [Zhou et al., 2016] Zhou, J. T., Pan, S. J., Tsang, I., and Ho, S.-S. (2016). Transfer learning for cross-language text categorization through active correspondences construction. In *AAAI Conference on Artificial Intelligence*.
- [Zhu et al., 2011] Zhu, Y., Chen, Y., Lu, Z., Pan, S. J., Xue, G.-R., Yu, Y., and Yang, Q. (2011). Heterogeneous transfer learning for image classification. In *AAAI Conference on Artificial Intelligence*.
- [Zoph et al., 2018] Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8697–8710.