Speech2Health: A Mobile Framework for Monitoring Dietary Composition from Spoken Data

Niloofar Hezarjaribi, *Student Member, IEEE,* Sepideh Mazrouee, *Student Member, IEEE,* Hassan Ghasemzadeh, *Senior Member, IEEE*

***Abstract*—Diet and physical activity are known as important lifestyle factors in self-management and prevention of many chronic diseases. Mobile sensors such as accelerometers have been used to measure physical activity or detect eating time. In many intervention studies, however, stringent monitoring of overall dietary composition and energy intake is needed. Currently, such a monitoring relies on self-reported data by either entering text or taking an image that represents food intake. These approaches suffer from limitations such as low adherence in technology adoption and time sensitivity to the diet intake context. In order to address these limitations, we introduce development and validation of Speech2Health, a voice-based mobile nutrition monitoring system that devises speech processing, natural language processing (NLP), and text mining techniques in a unified platform to facilitate nutrition monitoring. After converting the spoken data to text, nutrition-specific data are identified within the text using an NLP-based approach that combines standard NLP with our introduced pattern mapping technique. We then develop a tiered matching algorithm to search the food name in our nutrition database and accurately compute calorie intake values. We evaluate Speech2Health using real data collected with 30 participants. Our experimental results show that Speech2Health achieves an accuracy of 92.2% in computing calorie intake. Furthermore, our user study demonstrates that Speech2Health achieves significantly higher scores on technology adoption metrics compared to text-based and image-based nutrition monitoring. Our research demonstrates that new sensor modalities such as voice can be used either standalone or as a complementary source of information to existing modalities to improve accuracy and acceptability of mobile health technologies for dietary composition monitoring.**

***Index Terms*—Nutrition Monitoring, Mobile Computing, Wearable Sensors, Smart Health, Natural Language Processing, Speech Recognition, String Matching.**

I. Introduction

The incidence of chronic diseases and medical costs associated with management of conditions such as diabetes, cardiovascular disease, cancer, and obesity continue to rise nationally and worldwide [1], [2]. Interventions that help with prevention and/or self-management of chronic diseases play a crucial role in reducing healthcare costs as well as mortality and morbidity rates. Diet and physical activity are critical lifestyle interventions for self-management and prevention of many chronic diseases [3]. Research in nutrition monitoring and physical activity aims to enhance lifestyle behaviors focused on controlling dietary habits and promoting physical activity and exercise. The ability to accurately monitor physical activity and nutrition is important in delivering timely and effective clinical interventions in both self-management and prevention of the disease.

Diet self-monitoring is one of the earliest approaches to nutrition assessment [4], [5]. This approach involves meal recalls and food frequency questionnaires [6]. One of the self-monitoring techniques is the use of pen/pencil and paper diaries [7], [8]. Although these approaches have shown some success in weight loss programs [9], detailed self-monitoring can be cumbersome, time-consuming, and erroneous. Accordingly, adherence to these methodologies is reported to be low [10]–[12]. Consequently, more objective methods of diet and physical activity assessment are warranted. While assessment of physical activity using wearable sensors is now a reality, diet monitoring still relies on users being involved in recording food intake data. Examples of such technologies include smartphone apps that allow users to enter diet data (e.g., food name, portion size, etc.) or those that require end-users to take an image of their dietary intake. These technologies have been successful in detecting nutrient-related information such as food type and eating time. In many intervention and/or metabolic studies, stringent monitoring of overall dietary composition and energy intake is needed. While text-based nutrition monitoring is time-consuming and burdensome, image-based approaches are time-sensitive and may pose privacy concerns. Thus, new technologies that address limitations of previous nutrition monitoring systems may enhance scalability and clinical utility of mobile nutrition monitoring.

Our goal in this paper is to develop and validate a framework for nutrition monitoring from spoken data. We utilize advancements in speech recognition in order to improve the ease of data entry for nutrition monitoring. Speech is the most natural form of human communication. The role of voice input has also been studied in human-machine communication research [13]. Prior research suggests that using voice input is more beneficial when user's hands and/or eyes are busy, keyboard of the device is small, user is disabled, spelling is important, and/or natural language interaction is preferred. Furthermore, entering text on mobile devices is time-consuming and error-prone compared to entering text on a full-sized keyboard. This is particularly important because smaller mobile devices such as smartwatches are becoming a common platform for interacting with mobile services. Although voice input would intuitively improve user experience, there are fundamental

N. Hezarjaribi and H. Ghasemzadeh are with the School of Electrical Engineering and Computer Science, Washington State University (WSU), Pullman, WA, 99164–2752, USA e-mail: {n.hezarjaribi, hassan.ghasemzadeh}@wsu.edu.

S. Mazrouee is with the Computer Science Department at the University of California Los Angeles (UCLA), Los Angeles, CA 90025, USA email: sepideh@cs.ucla.edu.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2709333, IEEE Journal of Biomedical and Health Informatics

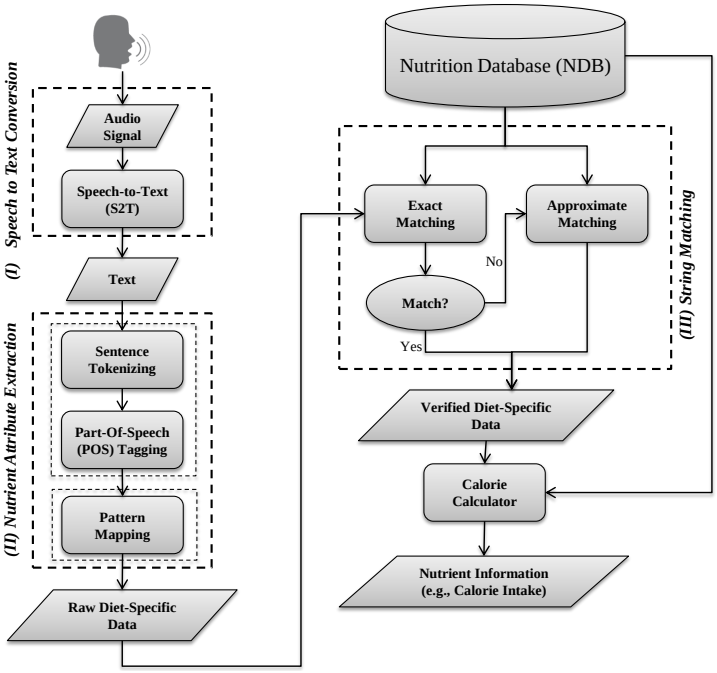

Fig. 1: Speech2Health Architecture: The system is comprised of three main modules including speech processing, natural language processing, and string matching each tailored towards nutrition monitoring.

challenges that need to be addressed due to the utilization of the system in uncontrolled end-user environments with background noise recorded during the input speech. Furthermore, voice input can generate unstructured text, which poses an additional level of complexity in understanding nutrition-specific data. Therefore, efficient natural language processing techniques are needed to identify nutrition-specific information such as food name and portion size. Moreover, text processing algorithms are needed to map the expressed food information onto dietary composition such as calorie intake.

Voice-based nutrition monitoring using mobile devices is a new and emerging research area. Our work in this paper builds based upon our recent work where we demonstrated the feasibility of using spoken language as a new modality for nutrition monitoring [14]. In this paper, we extend our prior work on multiple fronts to (1) improve the accuracy performance of our nutrient information extraction by developing a novel pattern mapping approach; (2) assess user acceptance of the proposed technology for nutrition monitoring in end-user settings based on data collected in naturalistic settings; (3) assess contribution of individual processing units to the overall accuracy of the system; and (4) examine robustness and generalizability of our framework using data collected in both noisy an noise-free environments. A high level overview of our approach is illustrated in Fig 1. Our contributions in this work can be summarized as follows.

- We introduce a new framework for voice-based nutrition monitoring and propose an array of data processing modules including a speech to text conversion, nutrient attribute extraction, and string matching to extract nutrient information from spoken language.
- We devise a *nutrient attribute extraction* by utilizing Natural Language Processing (NLP) and developing a novel pattern mapping approach for tagging nutrient data. To this end, we utilize NLP for tokenizing the sentence and tagging the words. Utilizing the tokenized words, their associated tags, and a dataset of extracted patterns, we assign nutrient-specific tags to the words (e.g., 'mac and cheese' $\rightarrow$ 'NN' + 'IN' + 'NN' $\rightarrow$ 'food name'). Our pattern mapping technique effectively identifies a variety of food names with various levels of complexity. This is an important new contribution because it demonstrates that we can utilize a given nutrition database to construct likely expected food patterns without collecting labeled training data from end users. As a result of this new development, we improve the performance of our calorie computation approach by 11.6% compared to our prior work [14].
- We develop a 2-tier string search algorithm including an exact matching and an edit-distance-based approximate matching to search food items from a given nutrition database (e.g., United States Department of Agriculture Food Composition Databases) and to compute nutrient values such as calorie intake, fat, and protein.
- The system is evaluated with real data collected with 20 subjects in an experimental setting that mimics noise-free as well as realistic noisy environments where the spoken data are entered into the system [14].
- We assess user acceptance of our nutrition monitoring approach with 10 additional participants to compare voice-based nutrition recording with two other diet recording systems including text-based and image-based methods in a week-long experiment.

II. Related Work

For a technology to be effective for behavior change, it needs to be persuasive [15]. Approximately 45% of American adults own a smartphone [16]; moreover, smartphones are

personal devices and are always-carried. Therefore, phones are ideal platforms for administrating eating behaviors as well as allowing users to monitor nutrition intake by providing a means for recording food and estimating calorie intake. Technological approaches to nutrition monitoring can be divided into three main categories; (1) wearable sensors; (2) computer vision; and (3) smartphone diaries.

The self-reported approaches (e.g., questionnaires) suffer from low user compliance and lack of sufficient accuracy due to the subjective nature of the monitoring [10], [11]. For the purpose of improving adherence and resolving the limitations of self-reporting, researchers have attempted to automate the task of nutrition monitoring [17]–[20]. A large spectrum of applications has been utilized for this purpose, ranging from mobile phone applications to on-body sensors [21]–[23] and computer vision [24], [25]. In the rest of this section, we briefly review the state-of-the-art for nutrition monitoring.

A. Wearable Sensors

Alshurafa et al. proposed a wearable system using piezoelectric sensors embedded in a necklace to detect food type based on skin motions [26]. This work distinguishes between solid and liquid foods. Sazonov et al. utilized piezoelectric sensors for chewing detection by capturing jaw movement [27]. The accuracy of eating moment and chewing detection was 85% and 81%, respectively. Thomaz et al. proposed a system for eating moment detection using accelerometer sensor embedded in a smartwatch [28]. The authors focused on eating style (e.g., 'with fork and knife', 'with hand', 'with a spoon'). Cheng et al. introduced a sensing system based on textile electrodes embedded in a neck collar wirelessly connected to a smartphone [29]. The technology focused on swallowing detection; hence, its utility is limited to portion control rather than estimating calorie intake. A major advantage of the aforementioned systems is that they perform nutrition monitoring seamlessly without the need for end-user to record nutrition data. However, such systems cannot be used to accurately estimate calorie intake, which is an essential component for diet-based interventions.

B. Computer Vision

He et al. proposed an automatic food type detection, called DietCam, based on computer vision by developing a multi-view multi-kernel Support Vector Machine (SVM) [24]. A dataset of 15262 web images was employed and divided into 55 classes. The accuracy of the classification algorithm ranged from 80% to 90% depending on the complexity of the food images. This system performs food type classification for popular American foods. One limitation of this approach is that, it does not provide information about food volume and calorie consumption. A food recognition system for Type 1 diabetic (T1D) patients based on bag-of-features model was proposed by Anthimopoulos et al. [25]. A dataset of 4868 images were collected from the web and categorized into 11 classes. The classification accuracy was 78%, on average. A leftover estimation system was proposed by Ciocca et al. for daily diet monitoring of canteen customers. The evaluation was conducted on 2000 images provided by 1000 participants. Images were classified into 11 Italian foods and the accuracy was 85% on average [30]. Chae et al. proposed a technique for volume estimation based on 3D reconstruction of images with an average error of 11% [31].One limitation of using computer vision for diet monitoring is the issue of time sensitivity. Image-based nature of this technique requires data recording prior to taking the food. However, it can increase the user's awareness negative eating habits [32]. Moreover, in some cases the visual information by itself may not be enough to accurately compute calorie intake.

C. Smartphone Diaries

There have been several studies on using mobile diaries for nutrition monitoring. Furthermore, a relatively large number of mobile applications for monitoring calorie intake and physical activity are publicly available. A prototype was created by Intille et al. called (POND) to motivate good dietary by providing just-in-time messages using bar-code scanner [33]. A mobile app, called Patient-Centered Assessment and Counseling Mobile Energy Balance (PmEB), was proposed for self-monitoring caloric balance by Tsai et al. [34]. The app consists of a client application on the mobile phone, a server application running on a web server, and a web-interface for registering and personalizing the mobile client. Another mobile app, called sapoFit, was presented by Rodrigues et al. [35], to keep track of daily personal health record (PHR). The app takes daily self-reported data provided by the user. This app keeps track of weight of the users based on the recorded information, calculates daily calorie intake, weekly weight loss, and alerts them in certain times. Although many available smartphone diaries provide the user with feedback about the nutrition intake, they require end-users to enter diet data manual on the app.

Algorithm 1 Nutrition Monitoring using Speech2Health

```
procedure SPEECH2HEALTH-ALGORITHM
Input: Audio waveform '𝒜𝒲'
Output: CalorieIntake
    Sentence set S = {s_1, s_2, ..., s_n} ← Speech-to-Text('𝒜𝒲')
    for each s_i in S do
        NutrientInfo ← NLPModule(s_i)
        FoodName ← NutrientInfo[0]
        PortionSize ← NutrientInfo[1]
        TimeStamp ← NutrientInfo[2]
        CaloriePerUnit←StringMatchingModule(FoodName)
        CalorieIntake←CalorieCalculator(PortionSize, CaloriePerUnit)
    end for
end procedure
```

III. SPEECH2HEALTH FRAMEWORK

An overall architecture of Speech2Health and major computational modules are shown in Fig 1. The goal is to compute nutrient information from spoken data that describes the food intake through natural language. In this section, we first present a high-level overview of the system. We then discuss individual components of the system including speech-to-text, nutrient attribute extraction, string matching, and nutrition database, in more details.

A. System Overview

As shown in Fig 1 and Algorithm 1, Speech2Health is a novel framework using speech recognition, text processing,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2709333, IEEE Journal of Biomedical and Health Informatics

and a 2-tier string search. In the proposed framework, the speech data is first converted into text in a *speech-to-text conversion* module. For this purpose, we use standard speech-to-text engines as discussed in Section III-B. The speech-to-text module will generate a set of words that will be further processed through our *nutrient attribute extraction* and *string matching*. Our hypothesis is that nutrition monitoring from spoken data can result in improving ease-of-use while achieving an acceptable accuracy performance in computing dietary composition such as calorie intake.

The nutrient attribute extraction text processing is responsible for sentence tokenization, POS-tagging, and complex food-name inference, as elaborated in Section III-C. We devise a new method of pattern mapping for performing diet-specific Part of Speech (POS) tagging on each token. The purpose of this module is to accurately identify food name and portion size numbers in the expressed sentence. As soon as nutrient-related information are located within the text, we feed the data into our *string matching* module, discussed in Section III-D. Finally, the *calorie calculator* module outputs calorie values based on the extracted data. This module extracts calorie values from the nutrition database based on the identified food information and a given nutrition bank that includes calorie intake and other nutrient information per serving of each food.

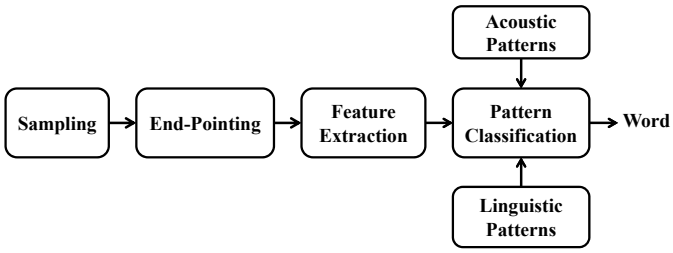

Fig. 2: Speech-to-Text (S2T) signal processing flow.

B. Speech-to-Text Conversion

Speech is the most natural form of human communication. Much research has been conducted on speech synthesizers, speech transmission systems, and automatic speech recognition. The input to such systems is a stream of sampled speech signals and desired output is a sequence of words. The audio signal is matched amongst existing patterns which represent different sounds.

A generic procedure for speech-to-text conversion is shown in Fig 2. There are four steps for converting voice to text: (1) Sampling: First step in 'Speech-to-Text' (S2T) is to sample the audio signal. The speech signal is sampled to generate discrete signal values and then samples are generated by digitizing the signal; (2) Endpointing: In this step, the presence of the speech is differentiated from non-speech regions to avoid additional computational complexity and prevent hallucination of unspoken words; (3) Feature Extraction: Pattern matching on the raw sample streams is not efficient. Therefore, some representations of input patterns are evaluated and important frequency domain features are extracted from the audio wave (e.g. energy, fundamental frequency, etc.); which refers to the analysis of signals with respect to frequency, rather than time. Feature extraction eliminates unrelated factors from input sequence. Speech is converted into a time-frequency representation, called spectrogram, in which phonemes have distinct dominant frequency bands; (4) Pattern Classification: The system is first trained using a set of known templates. The system is then used to match the spoken word with the trained templates by choosing the most likely template as the classified pattern.

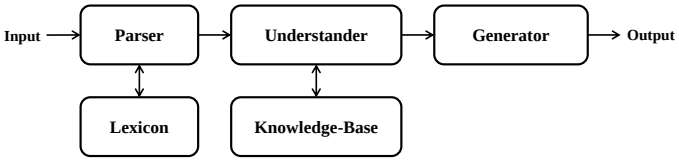

Fig. 3: General semantic diagram of the Natural Language Processing unit

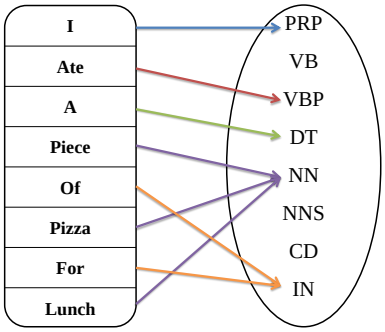

Fig. 4: An example of POS tagging. A set of words (on left) are labeled by the POS tagging as shown on the right.

C. Nutrient Attribute Extraction

The goal of our *nutrient attribute extraction* module is to learn a simple ontology for nutrition monitoring and to use this ontology to extract nutrition-related information from natural language text. In order to create such an ontology, we need to extract nutrition-specific terms and identify relationships between those concepts. To this end, Speech2Health leverages the nutrition database (i.e., USDA nutrition database) that was discussed previously to extract noun phrases from plain text using a linguistic processor and to encode them for automatic and easy retrieval during nutrition monitoring. We first use standard Natural Language Processing (NLP) algorithms to extract noun phrases associated with each food entry in the nutrition database. We run Part-of-Speech (POS) tagging on the nutrition database and obtain a set called Nutrient Pattern Set (NPS). Each element in an NPS is a pair of simple or complex tag and likelihood of the tag appearing in the nutrition database. These food-patterns are then sorted from the longest to shortest. For example, the food name *'mac and cheese'* encoded with the tag *NN+CC+NN* is longer than *'cheeseburger'* that is encoded with the tag *NN*. Based on this pattern set, we develop a pattern mapping algorithm to identify nutrition attribute from natural language text. These procedures are further discussed in the following subsections.

1) Natural Language Processing (NLP): NLP is the ability of a machine to understand humans' pronouncement as it is spoken. In the context of nutrition monitoring, we use a spoken language to understand diet intake of the user. Prior to extracting nutrition-specific concepts, we apply NLP for sentence tokenization. As shown in Fig 3, there are several major units in the NLP unit including parser, analyzes the input sentence syntactically; lexicon, is a dictionary of recognizable words; understander, performs semantic analysis in conjunction with the knowledge base; knowledge-base, the unit that contains structured and unstructured information required for text processing; and generator, usable outputs, also called tags.

In this paper, we use NLP for sentence tokenization and Part-of-Speech (POS) tagging. This is a supervised learning problem. An array of text-data is given to be labeled by a set of tags, as shown in Fig 4, where the words on the left are mapped to the tags shown on the right using the NLP algorithm. Each tag is assigned to a word based on the grammatical roles of the data.

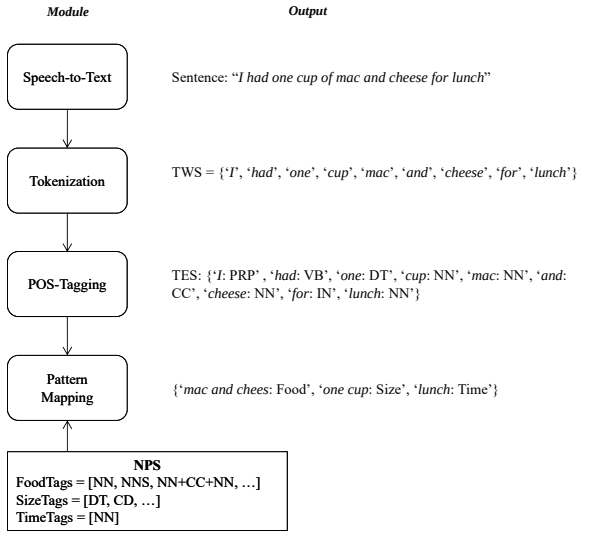

Fig. 5: An example of Nutrient Attribute Extraction; (1) input is a string; tokenization outputs a list of strings called Tokenized Words Set (TWS); (2) POS_tagging generates a list of tokens and their associated tags; (3) Pattern mapping takes a list of Nutrient Pattern Set (NPS) and attempts to find a pattern in TES and pass then on to String Matching module to check for availability; (4) output is a list of Tagged Nutrient Data

2) Pattern Mapping: The labels generated by the NLP unit are associated with individual words in the sentence. In reality, however, we need to put these tags in the context of nutrition monitoring to identify food names, portion sizes, and other nutrition-related information that one may report during diet recording. Our pattern mapping algorithm builds based on the outputs generated by the NLP module. After tokenizing and tagging the sentence and words in the NLP unit, we retrieve two sets including Tokenized Words Set (TWS) and Tags Equivalent Set (TES). In order to extract nutrient tags in each sentence, an adaptive window is utilized. The size of this window is initially set to the longest pattern according to the Nutrient Pattern Set (NPS) that is computed from the nutrition database. When a match is found among the patterns, the data is sent to the string matching module. If the data is found in the nutrition database, the window is moved to examine the rest of the sentence. Otherwise, the window size is reduced by one and the search process repeats. An example of the nutrient attribute extraction unit is shown in Fig 5 with an implementation of the algorithm described in Algorithm 2.

Algorithm 2 Pattern Mapping Algorithm

procedure PATTERN MAPPING
Input: Tags Equivalent Set, TES={ $t_1, \ldots, t_n$}
Input: Tokenized Words Set, TWS={ $w_1, \ldots, w_n$}
Input: Nutrients Pattern Set, NPS={$< T_m, p_m >, \ldots, < T_1, p_1 >$}
Output: Tagged nutrient data, {$< W_1, NT_1 >, \ldots, < W_m, NT_m >$}

    winSize←pattern with maximum length
    stepSize←winSize
    startingPoint←0
    initialize NutTag to NULL
    initialize ComplexWord to NULL
    while tag in TES **do**
        for i in range(startingPoint, startingPoint+winSize) **do**
            NutTag += t_i
            ComplexWord += w_i
            if NutTag matches a pattern in NPS **then**
                $W_{startingPoint}$←ComplexWord
                $NT_{startingPoint}$←NutTag
                availabilityFlag←StringMatching(NutTag)
                if availabilityFlag is True **then**
                    startingPoint += stepSize
                    break
                else
                    winSize←winSize - 1
                    StepSize←StepSize -1
                end if
            else
                winSize←winSize - 1
                StepSize←StepSize -1
            end if
        end for
    end while
end procedure

In order to locate food names, a predefined set of patterns, stored in Nutrient Pattern Set (NPS), are utilized. These patterns are defined by different combinations of tags which can form food names. For example, the food name *mac and cheese* comprises a pattern of form *'NN + CC + NN'*. As another example, *orange juice* is associated with the tag pattern *'NN + NN'*. Our food-name allocation maps the words in the expressed sentence onto meaningful patterns from a predefined pattern set. Later, the chosen words based on patterns can be processed for finding nutrition specific data.

To locate the portion size in a given sentence, Speech2Health first attempts to use existing NLP tags that infer portion size (such as CD (Cardinal Numbers) and DT (Determiner)). If a portion size is not found using existing tags, we utilize a predefined pattern set to map special keywords to sizes (e.g., (a: 1), (an: 1), and (half: 0.5)). In case the system cannot find any appropriate portion size values within the second step, a default portion size value is used for each food name.

The time-stamps may be expressed by the user. Speech2Health examines the sentence and looks for a set of keywords (e.g., breakfast, lunch, and dinner). If the keywords were not found in the sentence, the time data

entry using the Speech-to-Text (S2T) module is inferred as time-stamp.

D. String Matching

String matching refers to a class of algorithms with the aim of finding location of a given pattern within a larger string or text. After tagging one or combinations of the words in the nutrient attribute extraction module, nutrient information (food names, portion sizes, and time stamps) are produced. The string matching module intends to match the food name with predefined food entries in the nutrition database. To this end, we need an effective search methodology. Because food names may not exactly match the nutrition database, due to the speech to text conversion errors (e.g., I had one *Walmart* (*walnut*) for my breakfast.). In this paper, a 2-tier string matching algorithm is developed. The approach includes an exact matching followed by an approximate matching method.

1) Exact Matching: Our algorithm attempts to find an exact match for the detected food name in the nutrition database. Assuming that the located food name is a pattern **P** identified by our pattern mapping algorithm discussed in Section III-C2, the exact matching problem tries to find all occurrences of **P** in the database. The food names identified through our pattern mapping algorithm are associated with longest tag sequence that corresponds to a food name in the database. This helps us avoid confusion during search. For instance, *bean and cheese burrito* contains 4 matches with the database including *bean*, *cheese*, *burrito*, and *bean and cheese burrito*. The pattern mapping algorithm infers *bean and cheese burrito* as a single food name, the exact matching algorithm uses this food name associated with the longest tag sequence to search the nutrition database.

2) Approximate Matching: If the algorithm does not find the food name through an exact matching process, an approximate matching approach (Levenshtein algorithm) is utilized. The similarity of the food name and each entry in the nutrition database is calculated based on the number of operations needed to convert the food name to the database entry. The operations used for conversion include insertion, deletion, and substitution. Let $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_m\}$ denote the strings A and B of length n and m respectively. The edit distance between A and B is computed by d_{mn} such that

$$d_{ij} = \begin{cases} d_{i-1,j-1}, & \text{if } a_j = b_i \\ min(\delta_1, \delta_2, \delta_3) & \text{Otherwise} \end{cases} \quad (1)$$

where

$$\delta_1 = d_{i-1,j} + w_{del}(b_i)$$
$$\delta_2 = d_{i,j-1} + w_{ins}(a_j)$$
$$\delta_3 = d_{i-1,j-1} + w_{sub}(a_j, b_i)$$

and functions $w_{del}(a)$ and $w_{ins}(a)$ denote costs of deletion and insertion of symbol a. Furthermore, $w_{sub}(a, b)$ represents the costs of substituting symbol a with symbol b in a string. We also note that this dynamic programming formulation works based on the following initial values:

$$d_{i0} = \sum_{k=1}^{i} w_{del}(b_k) \qquad 1 \le i \le m$$

$$d_{0j} = \sum_{k=1}^{j} w_{ins}(a_k) \qquad 1 \le j \le n$$

We assign a unit cost (i.e., $w_{del}(a) = w_{ins}(a) = w_{sub}(a, b) = 1$) to each of the operations, Thus, we actually compute the the minimum number of operations required to transform string a to string b. The time complexity of this algorithm is θ(mn); therefore, performing a search on the entire database can be time-consuming. To minimize the amount of time complexity of the approximate matching algorithm, we perform it in two steps. In the first step, the algorithm aims to narrow down the search space in the nutrition database. This will generate a list of likely matches for the food name under search. In order to accomplish this goal, we define a similarity likelihood for each pair of food name in the database and the detected food name in the stated sentence. It is computed by dividing the number of matched characters by the total number of characters in the two strings. This probability is later compared against a predefined threshold; thereafter, the food names with the probabilities less than the predefined threshold will be eliminated from the search. In the second step, the food name most similar to the located word in the sentence is identified by running the edit distance algorithm on the probable match list obtained from the first step. The string matching procedure is shown in Algorithm 3.

Algorithm 3 String Matching module

```
1: procedure StringMatching
2: Input: FoodName, simTh
3: Output: CaloriePerUnit
4:     minDistance←+∞
5:     searchFlag←searchDB(FoodName)
6:     if searchFlag != NULL then
7:         CaloriePerUnit←DB(foodName)
8:     else
9:         for each element in DB do
10:            SimProb←sequenceMatcher(FoodName,element)
11:            if SimProb >= simTh then
12:                Distance←Levenshtein(FoodName,element)
13:                if Distance <= minDistance then
14:                    minDistance←Distance
15:                    minDistanceInd←index(element)
16:                end if
17:            end if
18:        end for
19:        FoodName←element with minimum distance
20:        CaloriePerUnit←DB(foodName)
21:    end if
22: end procedure
```

IV. Validation Approach

Our goal was to assess the performance of individual modules in Speech2Health as well as the performance of the entire system through a number of experiments and user studies. In this section, we explain implementation of the system and each experiment.

A. System Setup

Google provides a number of speech features for Android. One of these features is a speech API (Application Program Interface) for application developers [36]. This API displays

a "Speak now" dialog and waits for user input. After user taps the button, it streams the audio to Google's server and API responds using 'RecognizerIntent'. This intent is a constant for supporting speech recognition through 'intent.Android.speech.RecognizerIntent'. 'Android.speech.RecognizerIntent.EXTRA_LANGUAGE_MODE' informs preferred speech model to the recognizer while performing 'ACTION_RECOGNIZE_SPEECH'. This way, the spoken data is converted into natural language text.

We implemented our nutrient attribute extraction and string matching modules in Python. For nutrient attribute extraction module, we employed Natural Language Toolkit (NLTK), a well-known platform, implemented in Python, for NLP. In order to perform approximate matching, we utilized Levenshtein function embedded in pylev library. Furthermore, the nutrition dataset was obtained from the USDA database [37] and was embedded in a hash table for a fast lookup with a time complexity of $O(1)$.

B. In-lab Experiments

Although parsing and discovering nutrient specific data from a string is an important task on its own, stress testing of the speech-to-text app has its own limitations. Therefore, we conducted three experiments, two in-lab experiments, and one in-field user study. We obtained Washington State University Institutional Review Board (IRB) approval for conducting these experiments.

Our in-lab study was conducted in two phases; the first phase was to determine common speech phrases that Speech2Health may encounter; the second phase was to evaluate the accuracy of the system.

In the first experiment, our goal was to identify common patterns users use to express their nutrition intake. This experiment could help us to develop a realistic experimental plan for in-lab data collection. In order to do this, 10 subjects aged between 18 and 30 were recruited to participate in this experiment. The subjects were asked to write down their nutrition intake for three days. Our finding was that people mostly express their eating in three ways; complete sentences (e.g., *I ate an apple for breakfast.*); describing a scenario (e.g., *I went to McDonald and ordered a large meal of big mac.*); brief expression (e.g., *1 apple*).

The second experiment was conducted to evaluate each module and validate the whole system. For this purpose, 10 subjects were recruited. Based on the results obtained from the first phase, a script was developed. The script includes 50 sentences that express the nutrition intake. In order to simulate the realistic environment, four different ambient noise settings including no noise, street, music, and movie sounds were considered. Each subject was asked to read the script four times in presence of different noises. Same source of noises with a fixed sound level was considered to obtain consistent results. The output of Speech to Text Conversion was transmitted to the server for further data analysis and performance evaluation.

C. In-Field Study

The purpose of this study is to test the system in uncontrolled environments and conduct a comparison with other nutrition monitoring methods. Three data collection methods including voice-based, image-based, and text-based were utilized. The voice-based method uses the Speech2Health system. In the image-based technique, the data was gathered by taking an image of the food, while the text-based method was conducted using the well-known myFitnessPal app. An Android phone was given to each subject with pre-installed software. For ground-truth labeling, we modified speech to text module in a way that allows input correction by the user. Later, these edits are utilized to determine the accuracy of the app in a real world setting. The reported information by subjects can be converted to text incorrectly due to several reasons; noise in environment, non-native speaker, mistake in pronunciation, etc.

The experiment was performed with 10 different participants in six consecutive days. The experiment required participants to use each method for two days (6 days total). They were asked to record every single food item that they ate for the entirety of the experiment. The participants were also set on different schedules in order to ensure that usability and frequency of use will not be affected by ordering of the three apps. For instance, the first participant used the three apps in this order: image-based, voice-based, and text-based; the second participant first utilized voice-based, then text-based, and finally image-based tool. After the data collection completed, user acceptance was evaluated using a questionnaire. The survey included questions in five categories, including attention, interaction, trust, impact, and ranking. Responses were mostly in the form of Likert scale ranging from 1 to 5. The survey was taken by 9 out of the 10 participants.

V. Experimental Results

In this section, the results of each experiment are presented. The mobile app was modified in order to save the output in a text file and send to a server for further analysis.

A. Accuracy of Speech2Health

A script was provided to the subjects incorporating variety of food names. In order to increase the nutrient calculation accuracy, a similarity algorithm was embedded in the system. The algorithm aimed to fix error of the application and compensate incompleteness of the database. Besides finding the most similar food name to the word that user expressed, the probability of similarity between the word in sentence and the food names in database was also computed. This probability was compared with a predefined threshold, which was given as input to the algorithm. The accuracy values were calculated for thresholds ranging from 0.7 to 0.99. Our results contain accuracy of the Speech2Health system with and without noise, the impact of pattern mapping and approximate matching on the system, and user acceptance of the technology.

After audio signal is converted to text, the output is fed into the Nutrient Attribute Extraction module. This module is responsible for tokenizing and parsing the sentence. After tagging the tokens utilizing POS, nutrition specific data are located based on predefined set of patterns obtained from the database. There are 180 distinct patterns in the database. The

TABLE I: The first fifteen frequent patterns occurred in the database

Pattern	Frequency
NN + NN	22.54%
NN	14.33%
NN + NNS	8.85%
NNS	7.16%
JJ + NN	6.57%
NN + NN + NN	6.32%
JJ + NN + NN	3.53%
JJ + NNS	3.43%
VBN + NNS	2.64%
VBN + NN	2.19%
JJ + NN + NNS	0.99%
NN + NN + NNS	0.94%
NNS + NN	0.84%

most frequent patterns are shown in Table I. The sentences are tokenized which breaks them to stream of words (e.g. "I had a cup of mac and cheese for lunch" $\rightarrow$ 'I', 'had', 'a', 'cup', 'of', 'mac', 'and', 'cheese', 'for', 'lunch'). Next, each word is read by the POS tagging module and predefined tags are assigned to them (e.g. [('I', 'PRP'), ('had', 'VBD'), ('a', 'DT'), ('cup', 'NN'), ('of', 'IN'), ('mac', 'NN'), ('and', 'CC'), ('cheese', 'NN'), ('for', 'IN'), ('lunch', 'NN'), ('.', '.')]). This module detects food name, portion size, and time-stamp based on this assignment and patterns which are predefined (e.g. 'DT': 'a': PortionSize pattern, 'NN'+ 'CC'+ 'NN': 'mac and cheese': food name pattern, 'NN': 'Lunch': time-stamp dictionary). The accuracy is high when the purpose of tagging is the same as what tagger is generated for.

In this work the tagger is used for nutrition monitoring; therefore, a combination of tags was considered as food tags (e.g. NN + NN, VBN + NN, and NN + IN + NN). The accuracy of calculated calorie using Nutrient Attribute Extraction in presence of error-free text is calculated using equation 2, wherein $Cal(calculated)$ and $Cal(Actual)$ are the amounts of the calculated by Speech2Health and actual calorie of the script respectively. The accuracy of the calorie calculation using the system is 97.69% on average.

Given that the focus of Speech2Health system is nutrition monitoring, the performance is also computed using nutrition specific data (e.g. *I had **a** cup of **milk** with **2** tablespoons of **cream cheese** for **breakfast***.). Further, the performance of the system was calculated using equations 3 and 4. $\mid NSW_{FN} \mid$ and $\mid NSW_{PS} \mid$ are the total number of food names and portion sizes in the script respectively. While $|\overline{NSW}_{FN}|$ and $|\overline{NSW}_{PS}|$ is the total number of incorrectly detected food names and portion sizes respectively using the Speech2Health. The results are shown in Fig 6. Starting from threshold 0.7 to about 0.75, a certain number of non-related data matching (e.g. "piece" as "spice") occurred. Similarly, for the thresholds of about 0.95 to 0.99 certain number of nutrition specific data were not detected anymore. Therefore, the accuracy values in graphs are relatively stable toward the both ends of the x-axis.

$$ACC_{Calculation} = (\frac{Cal(calculated) - Cal(Actual)}{Cal(Actual)}) \times 100 \quad (2)$$

$$ACC_{FN} = (\frac{\mid NSW_{FN} \mid - \mid \overline{NSW_{FN}} \mid}{\mid NSW_{FN} \mid}) \times 100 \quad (3)$$

$$ACC_{PS} = (\frac{\mid NSW_{PS} \mid - \mid \overline{NSW_{PS}} \mid}{\mid NSW_{PS} \mid}) \times 100 \quad (4)$$

B. User Acceptance of Speech2Health

Our experiments for assessing user acceptance of the technology focused on five categories including attention, interaction, trust, impact, and rating. The utility of these in assessing user acceptance of technology has been demonstrated by other researchers in the past [38]–[40]. Responses were either on a Likert scale from 1 to 5 or a ranking of the three technology-based nutrition monitoring approaches. Table II shows our findings from this study.

Attention: The participants used voice-based food recording more often compared to the two other approaches. As shown in Table II, Speech2Health received a score of 3.89 out of 5. The participants were asked to evaluate each method on a scale of 1 to 5 where 1 denoted 'very infrequently' and 5 represented 'very frequently'. On average, participants used Speech2Health 7 times per day and the weighted average was 3.89. The food recording using image-based method had the least number of accesses. Moreover, the data using image-based showed that most of the participants remembered recording the food intake after they started eating. On average voice-based system was utilized 24.28% and 41.45% more than text-based and image-based respectively.

Interaction: As shown in Table II, one question was asked in this category. The question focused on the ease-of-use for each food recording method. The participants found the voice-based method easier to use with a score of 4.25 out of 5 compared to 3.13 for text-based approach and 2.25 for the image-based method. Voice-based was perceived to be 35.8% and 88.9% easier to use than text-based an image-based respectively.

Privacy: This question was designed to assess the awareness of privacy measure for each method. The participants felt that the text-based method provided more privacy with a score of 3.75 out of 5 compared to the two other approaches. From the comments provided by the participants, we noticed that voice-based users felt uncomfortable to record their food in public; likewise, users were not willing to take an image of their food due to the appearance of their dish in the taken image. On average, the privacy of the voice-based approach is 15.07% less than text-based and 49.58% higher than image-based methods.

Impact: This question aimed to perform a comparative assessment of the three methods in terms of their overall acceptance. Thus, the participants were asked to rank the three methods. Based on the collected survey data, 75% of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2709333, IEEE Journal of Biomedical and Health Informatics

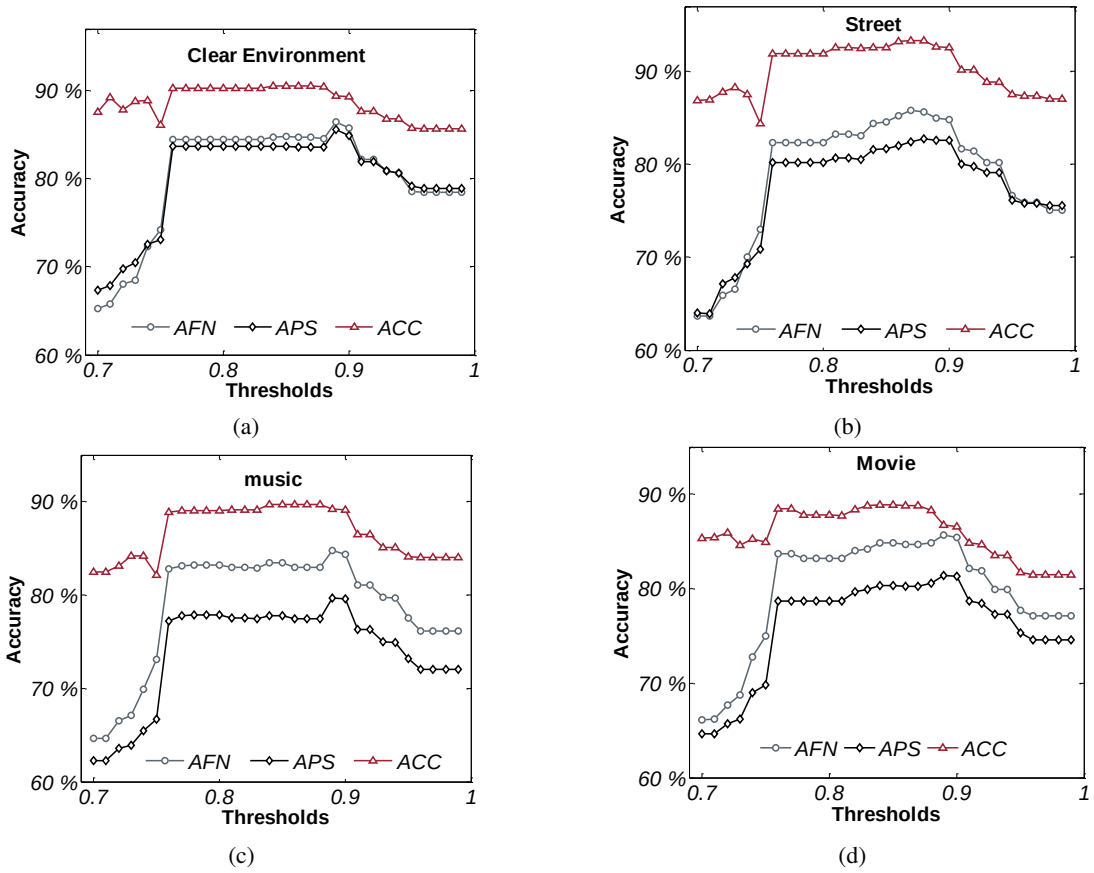

Fig. 6: Accuracy of the number of correctly detected food names (AFN), portion sizes (APS), and correctly calculated calorie (ACC) for a) clear environment, b) street, c) music, and d) movie.

TABLE II: Use acceptance results comparing speech-based, text-based, and image-based food intake recording.

Category	Speech	Text	Image
Attention: How frequently did you use the mobile app (1:very infrequent, 5:very frequent)?	3.89	3.13	2.75
Interaction: How satisfied are you with the app's ease of use (1:very dissatisfied, 5:very satisfied)?	4.25	3.13	2.25
Privacy: How satisfied are you with the privacy of using the apps (1:very dissatisfied, 5:very satisfied)?	3.56	3.75	2.38
Impact: What number would you use to rate each mobile app (1:best, 3:worst)?	2.44	2.00	1.56
Rating: How likely are you going to use Speech2Health as your nutrition monitoring app (1:very unlikely, 5:very likely)?	3.25	N/A	NA

the participants considered voice-based to be the best of the three technologies, 25% identified text-based as being the best, and none of the participants reported image-based as the best method. On a scale of 1-3 where 3 was the best approach and 1 denoted the worst method, voice-based, text-based, and image-based received weighted averages of 2.44, 2.00, and 1.56 respectively. Rate of using voice-based system was observed to be 22% and 56.41% more than text-based and image-based respectively.

Rating: The participants were asked to rank voice-based (Speech2Health) alone on a scale of 1–5. Our goal was to assess the likelihood of Speech2Health to be utilized as nutrition monitoring tool in the future. Among the 10 participants in our in-the-filed study, 2 implied they are going to use the app 'very likely', 2 expressed 'somewhat likely', and 4 chose 'somewhat unlikely'.

In addition to the above-mentioned quantitative measures, we also asked the participants to provide any other additional comments that they may have. Several thoughts emerged from participants' comments as follows.

The current version of Speech2Health misidentified the words expressed by fast speakers as well as some non-native speakers. This suggests that there is room for conducting research in this area to develop a more robust and comprehensive speech-to-text module in our framework.

The text-based nutrition monitoring method requires an instantaneous Internet connection for the app to be a usefulness tool. We quote this from one participant: *myFitnessPal has a limited local database, requires constant connection to the Internet to be useful, though the bar code search function is useful in some cases.*

A hybrid nutrition monitoring platform may be more advantageous as suggested by a participant: *Incorporating features from myFitnessPal (large database of foods, easy searching) and maybe an optional image component would give more avenues to get data imputed.*

An image-based nutrition monitoring was identified frustrating by several participants: *Taking pictures of every single*

little thing gets annoying after a while and *I don't think anyone would use the image-based app!!*.

We utilized several statistical tests to validate our user acceptance testing. First, we performed a t-test to assess whether there is a statistically significant difference between our method and each one of the other methods. Second, systematic difference between our voice-based method (gold standard) and two other methods (text-based and image-based) was assessed using a one-way Analysis of Variance (ANOVA) test. The goal of this ANOVA test was to identify any significant differences between user acceptance of our method and that of other methods.

The t-test analysis revealed that there is a significant difference between user acceptance of voice-based and image-based ($p < 0.03$) demonstrating that our method achieves a statistically significant higher user acceptance compared to the image-based nutrition monitoring. This test also showed the superiority of our method over the text-based method is not statistically significant thus suggesting the need for a larger scale user study.

Fig 7 depicts the results of the ANOVA tests (95% CI) on the user acceptance test. There was no significant difference between text based and our method; still, our method performed better. Image-based method had a significant difference from our method at the 0.95 significance level (p<.05).

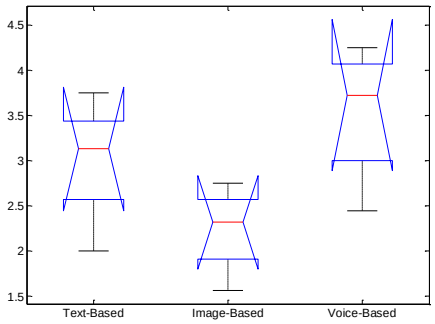

Fig. 7: The result of the ANOVA tests on the user acceptance test

C. Impact of Pattern Mapping

In order to locate the main source of error in the system we decided to evaluate performance of the Speech-2-Text Conversion. As mentioned previously, a script was provided to the subjects to read in presence of different noises that generates an audio signal for each environment. This audio signal is converted to text via the S2T app.

The performance of the system was calculated using equations 3 and 4 in absence of other modules (Nutrient Attribute Extraction and String Matching). The results are shown in Table III, wherein each number represents the accuracy of the system only using the S2T app in different environments. These numbers themselves do not represent the accuracy of the entire system. As explained previously, subsequent modules (Nutrient Attribute Extraction and String Matching) improve the accuracy of the system significantly.

TABLE III: Performance S2T conversion in various noise settings.

	clearEnv	**Street**	**Music**	**Movie**
ACC_{FN}	80.69%	77.31%	78.61%	77.40%
ACC_{PS}	80.82%	76.25%	76.33%	76.23%
ACC_{Total}	81.99%	77.17%	77.49%	76.89%

D. Impact of Approximate Matching

After detecting the food names and their corresponding portion sizes, calorie value was calculated for the script. Since speech-to-text conversion may generate noisy output, some of the food names were either not detected or incorrect. Moreover, the approximate matching tries to correlate non-nutrition specific data with the food names in the database.

The string matching module contains two matching approaches including an exact matching and an approximation matching as discussed before. In the exact matching, the algorithm checks for the word detected as food name in the database and finds an exact match for the detected word. If the word was not found in the database, it tries to find the most similar food name to that. Approximation matching is implemented for improving the performance of speech recognition module. In order to make the process fast, given that all detected words with Nutrient Attribute Extraction module are not nutrition specific data, we first set a threshold of similarity on our search and narrow down the database using Sequence Matcher provided by difflib library. This module tries to find the longest matching block between two words. Then we use edit distance to fix the mispronunciation of the app.

The program was tested with different threshold values ranging from 0.7 to 0.99. The lower thresholds resulted in non-related data matching (e.g. "piece" as "spice"). On the other hand, by increasing the threshold beyond an optimal value, some of the nutrition specific data were not detected anymore (e.g. "Apple", "Apples"). We experimentally find the optimal value of the threshold. The results are shown in Fig. 6 which implies that optimal threshold value is around 0.85. The accuracy of the system is 92.2% for noise-free and 90.4% for noisy environment, respectively.

Fig 8 shows the performance of the Speech2Health with and without approximation matching. The results demonstrate that the amount of improvement in finding nutrition- specific data using approximate matching is 2.01% for noise-free and 6.1% for noisy environments.

E. Comparison

As stated previously, nutrition monitoring from spoken data is a new research area. To the best of our knowledge, the work by Korpusik et al. [41] is the only prior research that attempted to use human speech data for nutrition monitoring. In this section, we discuss main differences between our work and the work in [41].

The primary focus on the work by Korpusik et al. is on language understanding aspects of that the system (i.e., semantic tagging the sentences expressed by users) which is also relevant to our approach for recognizing food patterns and searching those patterns in a given nutrition database such

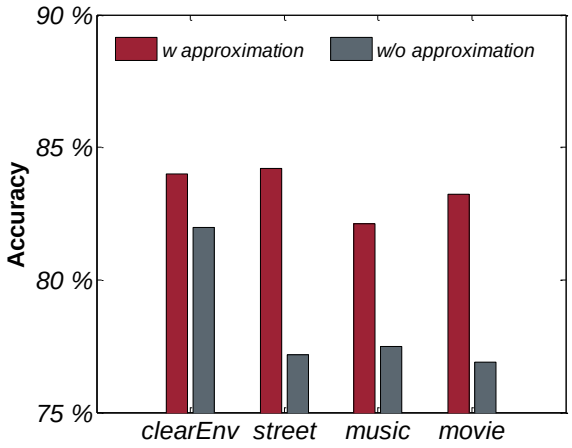

Fig. 8: Performance of finding nutrition-specific data with and without approximate matching algorithm

as USDA's. For this purpose, authors develop a Conditional Random Field (CRF) model to learn common ways meal descriptions are expressed by the users. As a result, the system requires a training phase during which labeled training instances (i.e., group truth meal descriptors) are collected and used to build the CRF model. In contrast, our work, Speech2Health, does not require any labeled training data for food pattern learning. The only source of knowledge in our framework is the nutrition database (i.e., USDA's database) which contains a comprehensive list of common food names and patterns. We, however, acknowledged the fact that learning food patterns from a set of labeled training data will potentially result in a higher accuracy for food pattern recognition. Yet, the recognized food patterns cannot be used for nutrient information extract (e.g., calorie intake computation) if those patterns are not found in the USDA database. While this learning aspect is the main difference between our approach and that of Korpusik et al., there exist several other differences as follows.

We implemented our system on mobile devices where all data collection and speech-to-text conversions occur in real-time. In contrast, authors in [41] utilize Amazon Mechanical Turk (AMT) for data collection and ground truth labeling. In particular, they build a speech recognizer from audio recordings of meal descriptions and label the recognizer's output on Amazon Mechanical Turk (AMT) to provide labeled data for CRF training.

Our evaluation in this paper consists of various test scenarios that mimic naturalistic settings. That is, we collect data in environments with various background noises. In contrast, previous research does not evaluate the performance of the food pattern recognition algorithm under different environmental noise.

Our work in this paper uses food pattern data for calorie intake estimation while previous research does not evaluate the utility of the identified food patterns for calorie intake computation. The accuracy of Speech2Health in calorie calculation is 92.2% in noise-free and 90.4% in noisy environments.

We acknowledge that the work by Korpusik et al. achieves a higher accuracy performance in food pattern identification potentially due to the training of a CRF model. We, however, note that not all those food patterns can be used for nutrition monitoring and specifically for calorie intake computation because only 71% of the identified food patterns are found in the USDA database. In contrast, more than 97% of the food items identified by our pattern mapping module match with an entry in the USDA database.

VI. Conclusions and Ongoing Research

The major contribution in this work is the introduction and validation of a novel approach for nutrition monitoring based on spoken data. This research provides a pervasive approach for recording and understanding spoken language for diet assessment by integrating advances in speech recognition, NLP, text analysis, and mobile health. The goal of using a speech-to-text app is to provide a more convenient nutrition monitoring tool for users in different situations. We utilized NLP algorithms to identify nutrition-specific information within the generated text. A 2-tier approach was devised for analyzing the text and calorie computation. The performance of the system in presence of error-free text is 97.7%. Although the performance of the speech-to-text app is 82.0% individually, the calorie calculation accuracies of the erroneous text using the system are 92.2% and 90.4% for noise-free and noisy environments respectively.

The accuracy of our system can not be compared with text and image-based approaches, due to the difference in modalities. Still, we performed a user acceptance analysis for comparing Speech2Health with two other nutrition monitoring techniques. The results show that Speech2Health achieves significantly higher scores in technology adoption metrics compared to the other approaches.

Our ongoing evaluation of Speech2Health involves in-the-wild study for using the platform in natural setting. Our goal is to measure the impact of the technology on clinical outcomes as well as assessing usability and acceptability of it. We are working on developing a prompting interface for improving the system. This interface is to allow end user enter a nonexistent food name, correcting the input, and providing missing information. Another goal is to provide the user with other nutrient values (such as protein, fiber, and lipid) other than calorie intake. Moreover, we are planning to use our system in conjunction with other technologies and develop a hybrid nutrition monitoring platform.

In this paper, we utilized the USDA dataset, which contains standard portion size for each food item. As part of our ongoing work, we are developing approaches that address the issue of portion size variation among individuals and over time. In particular, we are developing two types of strategies: (1) algorithms that identify nutrition behavior of the user and utilize user's behavior to automatically identify portion sizes; (2) developing prompting technologies that allow for interaction with users to enter their portion size in real-time. Taking care of this requires a thorough interaction between the user and the system and prompting multiple times to get the exact size from them. Most of the times, users only response to few number of prompts and get frustrated by high numbers

of them. One solution can be reporting weight instead of other units of measurement. But many users prefer reporting units such as piece, slice, glass instead of pound, ounce, and gram. We appreciate your comment. We agree that a single portion size can vary from one individual to another. Considering equal size among all the users may cause an error in nutrient calculation.

REFERENCES

[1] R. W. Kimokoti and B. E. Millen, "Diet, the global obesity epidemic, and prevention," *Journal of the American Dietetic Association*, vol. 111, no. 8, pp. 1137–1140, 2011.

[2] E. A. Finkelstein, I. C. Fiebelkorn, G. Wang *et al.*, "National medical spending attributable to overweight and obesity: how much, and who's paying?" *Health affairs-millwood va then bethesda ma*, vol. 22, no. 3; SUPP, pp. W3–219, 2003.

[3] M. Nestle, *Food politics: How the food industry influences nutrition and health*. Univ of California Press, 2013, vol. 3.

[4] R. R. Wing and J. O. Hill, "Successful weight loss maintenance," *Annual review of nutrition*, vol. 21, no. 1, pp. 323–341, 2001.

[5] B. B. Qi and K. E. Dennis, "The adoption of eating behaviors conducive to weight loss," *Eating behaviors*, vol. 1, no. 1, pp. 23–31, 2000.

[6] L. R. Wilkens and J. Lee, *Nutritional epidemiology*. Wiley Online Library, 1998.

[7] L. E. Burke, S. M. Sereika, E. Music, M. Warziski, M. A. Styn, and A. Stone, "Using instrumented paper diaries to document self-monitoring patterns in weight loss," *Contemporary clinical trials*, vol. 29, no. 2, pp. 182–193, 2008.

[8] R. L. Collins, T. B. Kashdan, and G. Gollnisch, "The feasibility of using cellular phones to collect ecological momentary assessment data: application to alcohol consumption." *Experimental and clinical psychopharmacology*, vol. 11, no. 1, p. 73, 2003.

[9] M. J. Devlin, "Obesity: Theory and therapy," *The Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 6, no. 3, pp. 325–327, 1994.

[10] K. B. Michels, "A renaissance for measurement error," *International journal of epidemiology*, vol. 30, no. 3, pp. 421–422, 2001.

[11] D. R. Jacobs Jr, "Challenges in research in nutritional epidemiology," in *Nutritional Health*. Springer, 2012, pp. 29–42.

[12] M. K. Mattfeldt-Beman, S. A. Corrigan, V. J. Stevens, C. P. Sugars, A. T DALCIN, M. J. Givi, and K. C. Copeland, "ParticipantsâĂŹ evaluation of a weight-loss program," *Journal of the American Dietetic Association*, vol. 99, no. 1, pp. 66–71, 1999.

[13] P. R. Cohen and S. L. Oviatt, "The role of voice input for human-machine communication," *proceedings of the National Academy of Sciences*, vol. 92, no. 22, pp. 9921–9927, 1995.

[14] N. Hezarjaribi, C. Reynolds, D. Miller, N. Chaytor, and H. Ghasemzadeh, "S2ni: A mobile platform for nutrition monitoring from spoken data," in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2016.

[15] B. J. Fogg, "Persuasive technology: using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, p. 5, 2002.

[16] L. Rainie, "Smartphone ownership update: September 2012," *Pew Internet & American Life Project*, 2012.

[17] Y. Dong, "Tracking wrist motion to detect and measure the eating intake of free-living humans," Ph.D. dissertation, Clemson University, 2012.

[18] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," *Biomedical and Health Informatics, IEEE Journal of*, vol. 18, no. 4, pp. 1253–1260, 2014.

[19] C. K. Martin, H. Han, S. M. Coulon, H. R. Allen, C. M. Champagne, and S. D. Anton, "A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method," *British Journal of Nutrition*, vol. 101, no. 03, pp. 446–456, 2009.

[20] K. Yatani and K. N. Truong, "Bodyscope: a wearable acoustic sensor for activity recognition," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 341–350.

[21] P. Alinia, R. Saeedi, R. Fallahzadeh, A. Rokni, and H. Ghasemzadeh, "A reliable and reconfigurable signal processing framework for estimation of metabolic equivalent of task in wearable sensors," *IEEE Journal of Selected Topics in Signal Processing*, 2016.

[22] H. Ghasemzadeh, R. Fallahzadeh, and R. Jafari, "A hardware-assisted energy-efficient processing model for activity recognition using wearables," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 21, 2016.

[23] N. Hezarjaribi, R. Fallahzadeh, and H. Ghasemzadeh, "A machine learning approach for medication adherence monitoring using body-worn sensors," in *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2016, pp. 842–845.

[24] H. He, F. Kong, and J. Tan, "Dietcam: Multi-view food recognition using a multi-kernel svm," 2015.

[25] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," *Biomedical and Health Informatics, IEEE Journal of*, vol. 18, no. 4, pp. 1261–1271, 2014.

[26] N. Alshurafa, H. Kalantarian, M. Pourhomayoun, J. J. Liu, S. Sarin, B. Shahbazi, and M. Sarrafzadeh, "Recognition of nutrition intake using time-frequency decomposition in a wearable necklace using a piezoelectric sensor," *Sensors Journal, IEEE*, vol. 15, no. 7, pp. 3909–3916, 2015.

[27] E. S. Sazonov and J. M. Fontana, "A sensor system for automatic detection of food intake through non-invasive monitoring of chewing," *IEEE sensors journal*, vol. 12, no. 5, pp. 1340–1348, 2012.

[28] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 1029–1040.

[29] J. Cheng, B. Zhou, K. Kunze, C. C. Rheinländer, S. Wille, N. Wehn, J. Weppner, and P. Lukowicz, "Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband," in *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 2013, pp. 155–158.

[30] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition and leftover estimation for daily diet monitoring," in *New Trends in Image Analysis and Processing–ICIAP 2015 Workshops*. Springer, 2015, pp. 334–341.

[31] J. Chae, I. Woo, S. Kim, R. Maciejewski, F. Zhu, E. J. Delp, C. J. Boushey, and D. S. Ebert, "Volume estimation using food specific shape templates in mobile image-based dietary assessment," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2011, pp. 78 730K–78 730K.

[32] L. Zepeda and D. Deal, "Think before you eat: photographic food diaries as intervention tools to change dietary decision making and attitudes," *International Journal of Consumer Studies*, vol. 32, no. 6, pp. 692–698, 2008.

[33] S. S. Intille, C. Kukla, R. Farzanfar, and W. Bakr, "Just-in-time technology to encourage incremental, dietary behavior change," in *AMIA Annual Symposium Proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 874.

[34] C. C. Tsai, G. Lee, F. Raab, G. J. Norman, T. Sohn, W. G. Griswold, and K. Patrick, "Usability and feasibility of pmeb: a mobile phone application for monitoring real time caloric balance," *Mobile networks and applications*, vol. 12, no. 2-3, pp. 173–184, 2007.

[35] J. J. Rodrigues, I. M. Lopes, B. M. Silva, and I. d. L. Torre, "A new mobile ubiquitous computing application to control obesity: Sapofit," *Informatics for Health and Social Care*, vol. 38, no. 1, pp. 37–53, 2013.

[36] B. Ballinger, C. Allauzen, A. Gruenstein, and J. Schalkwyk, "On-demand language model interpolation for mobile speech input." in *INTERSPEECH*, 2010, pp. 1812–1815.

[37] United states department of agricalture (usda). [Online]. Available: http://www.usda.gov/wps/portal/usda/usdahome

[38] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, pp. 425–478, 2003.

[39] H.-P. Shih, "An empirical study on predicting user acceptance of e-shopping on the web," *Information & Management*, vol. 41, no. 3, pp. 351–368, 2004.

[40] Y.-S. Wang, Y.-M. Wang, H.-H. Lin, and T.-I. Tang, "Determinants of user acceptance of internet banking: an empirical study," *International journal of service industry management*, vol. 14, no. 5, pp. 501–519, 2003.

[41] M. Korpusik, C. Huang, M. Price, and J. R. Glass, "Distributional semantics for understanding spoken meal descriptions," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 2016, pp. 6070–6074. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2016.7472843