IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 10, 2022, accepted March 24, 2022, date of publication April 4, 2022, date of current version April 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3164419

Speech Recording for Dietary Assessment: A Systematic Literature Review

CONNOR T. DODD[1], MARC T. P. ADAM[1], AND MEGAN E. ROLLO[2]
[1]College of Engineering, Science and Environment, University of Newcastle, Newcastle, NSW 2308, Australia
[2]Faculty of Health Sciences, Curtin University, Bentley, WA 6102, Australia

Corresponding author: Connor T. Dodd (connor.dodd@newcastle.edu.au)

This work was supported in part by the Bill and Melinda Gates Foundation under Grant OPP1171389, and in part by the Australian Government Research Training (RTP) Scholarship.

ABSTRACT Traditional methods of capturing people's dietary intake are complex and labour-intensive, requiring a high level of literacy and time. Speech recording has potential to reduce these barriers, and recent technological advances have greatly increased the viability of this approach. The aim of this paper is to establish the current state of research on the usage of speech records in dietary assessment. To this end, we performed a systematic literature review and summarised the current state of research along a conceptual framework that captures the components involved in using speech records for dietary assessment. Six databases from the nutrition and computing domains were interrogated, resulting in 21 relevant papers. Speech recording in an unstructured format was preferred when compared against other methods by all three studies reporting comparisons. High technological satisfaction and ease of use were noted by all eight studies reporting user acceptance. When recording data, 78% of studies focused on collecting prospective food records. The choice of device reflected this, with 15 of 18 studies reporting a form of handheld, portable collection device intended to be always available. To process data, nine studies performed automated speech transcription achieving an average accuracy of 83%, seven of which utilized a readily available commercial service. Of the five studies that used natural language processing to further automate analysis, an average accuracy of 82% was reported. Further research is required to adapt these prototypes to address practical challenges in dietary assessment and monitoring (e.g. self-monitoring for low-literacy users).

INDEX TERMS Automatic speech recognition, dietary assessment, food recording, natural language processing, systematic literature review.

I. INTRODUCTION

Sub-optimal diet quality is the primary risk factor for many of the leading causes of chronic disease and death, including diabetes, cardiovascular disease, obesity, and cancer [1]. To improve diet quality the current diet must be analysed through a process of collection and interpretation called dietary assessment. There is a variety of collection methods utilised for different situations and populations, but they are widely complex and labour intensive [2], and in populations with varying levels of literacy traditional methods such as self-administered written records are not viable [3].

Speech recording has been explored as a tool to collect intake descriptions in varying-literacy populations for some time, as it allows for data collection where literacy levels would affect traditional methods [4], [5]. Speech is also a natural and quick method of communication [6] and may reduce participant burden resulting in more complete records. However, speech has traditionally been expensive to collect and parse, requiring a large amount of costly manual work by professional staff. Recent advances in speech recognition technology have greatly reduced this cost, providing the ability to automate a portion of this work with increasing accuracy [7], [8]. Widespread adoption of smartphones capable of recording and processing raw speech now provide an accessible conduit for data collection. Speech input is also gaining consumer acceptability as commercial products with speech interaction are introduced (e.g. Amazon Alexa, Apple Siri) [16].

Due to these advances, there has been an increased research interest in using speech to perform dietary assessment. With a global shift toward improving diet-related health outcomes such as obesity and malnutrition in low-income populations [9], and the potential of new speech recognition tools being

The associate editor coordinating the review of this manuscript and approving it for publication was Giovanni Dimauro.

realised, there is a need for a comprehensive overview of the current state of the literature. In this paper, we perform a systematic literature review (SLR) with the following aims: To summarise existing literature surrounding the use of speech recording to collect dietary intake; to provide an overview of the goal of collecting speech-based dietary information, and the associated issues; and to evaluate the state of the literature and identify paths for further research. By meeting these aims, this review will serve to address the current gap in the literature and better position future research in this area.

The following chapter is structured as follows: section II provides a background of the topic area and distinction of common terms. Section III describes the method used to perform the SLR, and section IV presents the conceptual framework. This is followed by section V presenting the results of the SLR. Finally, section VI discusses the results and presents paths for further research.

II. BACKGROUND

A. DIETARY ASSESSMENT AND SELF-MONITORING

Dietary intake assessment is essential for surveilling population nutritional adequacy, identifying relationships between dietary intake and health outcomes, and to determine the effect of nutrition interventions. In addition, chronic conditions, such as diabetes, rely on dietary self-monitoring as a key management strategy [9], [10]. Food records are one method which can be used to assess intake and to self-monitor intake [11]. In this method, each food or drink a person consumes is recorded in some manner, usually with a quantity and enough detail to identify it later. Traditionally this was performed through pen-and-paper diaries, which a person could maintain for a specified time period and provide to a dietitian or researcher for analysis producing an estimate of food and nutrient intake. Today, this process is commonly completed through a smartphone or web app. This technology allows for faster, more complex approaches, including additional data such as images or audio for more accurate identification. This data can then be provided to dietitians for analysis, but some commercial approaches make use of the device to process the data into immediate feedback for the user. These methods require more user interaction, such as recording a barcode or selecting from a list of foods [31].

One can distinguish *prospective* and *retrospective* methods for dietary intake assessment. Food records are considered a prospective method with data captured at the time of consumption. This allows for more complete data capture, but in practice the burden of performing this recording leads to significant reactivity bias [12]. In contrast, the 24hr recalls and food frequency questionnaires are retrospective methods. In these methods, intake data are collected after eating with reference time periods ranging from the previous day for 24hr recalls through to days, weeks, months or years for food frequency questionnaires, depending on the actual tool used. Retrospective methods are more prone to errors relating to memory [11]. Choosing a collection method is often a trade-off between collecting sufficient detail and not overburdening participants and researchers with time costs. Accurate collection and complexity are closely linked, wherein the difficulty the participant experiences describing food items in sufficient detail directly affects the accuracy the researcher can obtain from the recording. Time cost affects both parties. Recording all food items consumed is a time-consuming task for participants, and combined with the high level of detail and high literacy skills required, this is a challenging activity to comply with [2]. The task of converting descriptions to nutrient intake data is also non-trivial, as each item must be identified and quantified, matched to a Food Composition Database (FCD) item, and the nutritional content calculated and analysed. For example, a simple description like ''a small orange'' requires the researcher to find the most appropriate item from the FCD (''Oranges, raw, all commercial varieties'' in the U.S. Dept of Agriculture database), consult measure references to estimate the quantity of ''small'', and then calculate the nutrient values for that quantity. This process is a significant time burden for researchers. One method of reducing this burden is automation through the use of speech recognition technologies.

B. SPEECH RECOGNITION

Speech recognition is a broad term that encompasses a range of processes used to extract content from audio recordings of human speech. The first process is to convert the audio into text, often called speech-to-text. Extraction of data from this text is called Natural Language Processing (NLP). NLP parses unstructured text such as transcribed speech in much the same way as a human would, identifying subjects and descriptors and linking them together to facilitate extraction [7]. Both speech-to-text and NLP technologies have recently seen great improvements as a result of machine learning advances and increased interest. There has also been a popular adoption as these technologies have become more commercially accessible, as seen in digital assistants such as Google Home and Siri.

These improvements are the driving force behind an increase in recent research on the topic. Speech recognition technologies are now performing at a high enough accuracy that automation of unstructured, raw audio data is now feasible [8]. As noted in the previous section, time cost can be a significant barrier to data collection and processing when measuring dietary intake. Collecting fast and simple audio records, and automatically processing these records, has the potential to make accurate dietary assessment more accessible. The viability of this approach is further supported by commercial developments, including the wide proliferation and accessibility of consumer-grade handheld devices and NLP software. Early attempts required researchers to construct and distribute custom collection devices [13], whereas modern studies are deployed to the participant's own smartphone [14]. Similarly, processing can now be accomplished with commercially available services such as Google Speech-to-Text for transcription [15] and SpaCy for NLP [16].

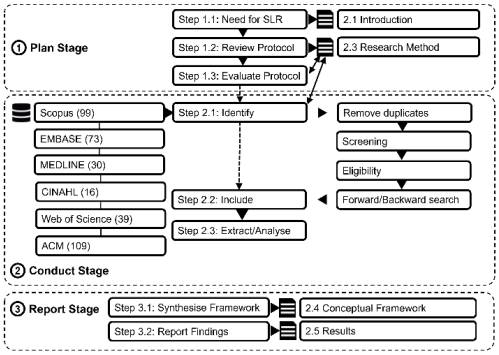


FIGURE 1. Stages of conducting the literature review.

III. RESEARCH METHODOLOGY

Dietary assessment is a health and nutrition domain, and the use of speech recordings as an input method originated here. While any application of dietary assessment is intrinsically linked to this domain, much of the recent literature has been published in computing journals. This research largely explores the processing of data rather than its application, being conducted in a different manner to health and nutrition. We chose to perform this SLR in a style popular with computer science and software engineering, as the recent advances in this area are what is motivating current research. As such, the systematic literature search follows software engineering guidelines [4]. To facilitate the comparison of literature across multiple disciplines, a conceptual framework was devised following the approach of Baumeister and Leary [17], a method used successfully for multidisciplinary computing and health literature reviews [18].

A. OVERVIEW

SLRs in this domain employ three phases: plan, conduct and report (see Figure 1). The planning phase begins with identifying the need for a SLR, which is clearly expressed in the introduction. An initial search was conducted across common databases and search engines to ensure this SLR was unique. A comprehensive search string and review protocol was also defined. During the conduct stage, the search is executed and filtered per the review protocol, and the results analysed.

B. SEARCH STRATEGY

During the initial exploratory search, several keywords were identified. Due to the cross-disciplinary nature of the research question and the evolution of the domain over time there is a variety of keywords with similar meanings. To ensure the search string was comprehensive, relevant papers were identified from the search and references and explored for further keywords. The structure of the search string is application AND modality.

The final search string is defined as: (''diet* assessment'' OR ''diet* intake'' OR ''diet* diary'' OR ''diet* record'' OR ''diet* recall'' OR ''nutri* assessment'' OR ''nutri* intake'' OR ''nutri* diary'' OR ''nutri* record'' OR ''nutri* monitoring'' OR ''food record'' OR ''food log*'' OR ''food journal'' OR ''food diary'' OR ''food recall'') AND (''voice record*'' OR ''voice transcri*'' OR ''natural language'' OR ''speech processing'' OR ''audio recording'' OR ''spoken language'' OR ''speech recognition'' OR ''voice recognition''). To cover the multiple disciplines, this search was executed on the literature databases: Scopus, EMBASE, MEDLINE, CINAHL, ACM and Web of Science.

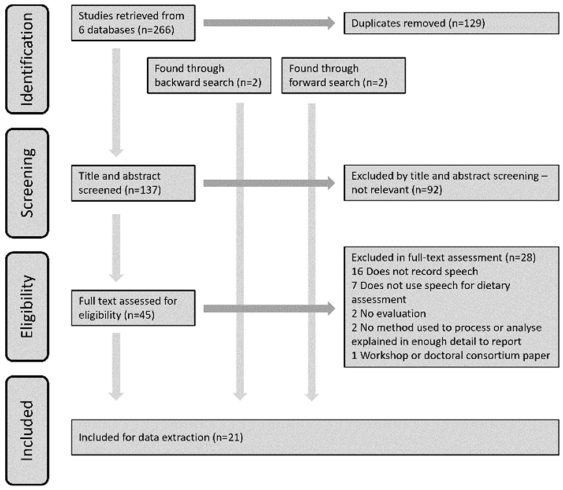


FIGURE 2. Flow diagram of article selection for the SLR.

Search results must be refined by implementing a review protocol, using inclusion and exclusion criteria. To be included the study must record human speech, to keep the focus on speech recording. This speech must also contain dietary information like food items. A quantitative evaluation must also be a component of the study. Studies were excluded if there was no method used to process or analyse dietary data from recorded speech explained in enough detail to report. Workshop and doctoral consortium papers were also excluded.

First duplicates were removed. Next the results of the search were interrogated in a title and abstract review to remove any studies that could be excluded based on the title and abstract alone. This was conducted by two authors using the inclusion and exclusion criteria, with a third author to resolve differing decisions. A full-text review was then run on the remaining studies in the same manner. Finally, forward and backward searches were conducted on included studies to identify any potential missing studies. A flow chart of applying these selection criteria can be seen in Figure 2.

IV. CONCEPTUAL FRAMEWORK

Given the interdisciplinary nature of the research at the intersection of computing and nutrition, we structure the literature review along a conceptual framework that incorporates the technical aspects of the approach as well as the study design and dietary assessment requirements. By abstracting and describing the different steps that must be undertaken to extract nutrient information from a speech description of food, the framework can compare the individual processes utilised by each study across the disciplines (see Figure 3).

FIGURE 3. Conceptual framework to structure the literature review.

A. STUDY DESIGN

The top level of the framework is the meta-design of the study. This involves the participants, setting and comparator. Participant characteristics heavily influence both the application and structure of the study, as specific demographics such as presence of chronic disease or varying literacy will have different reasons and requirements for obtaining accurate data. The setting in which data are collected also informs study design, where free living studies will have vastly different requirements to studies in a controlled or semi-controlled environment. Finally, one or more comparators are often collected alongside the speech recordings to allow for evaluation. These may be another method of recording intake to compare with speech recordings, or data used to evaluate the efficacy of speech in some other capacity, such as user acceptance questionnaires.

B. DATA RECORDING

1) CAPTURING DEVICE AND TECHNOLOGY

In the framework, we separate the device used to capture speech and the technology to process it to allow for the distinction between hardware and software. Thereby, a device would be a physical entity like a smartphone, the technology would be the application running on the smartphone used to record speech.

2) CAPTURE TYPE

The type of capture is the way participants are instructed to capture data relating to their intake. The primary methods would be prospective and retrospective, discussed in Background section A. Prospective capture refers to the creation of food records made at the time of consumption. Retrospective capture broadly refers to food data collected after consumption but is generally further defined with a specific method such as 24hr recall or diet history.

3) COMPLEMENTARY RECORDINGS

Complementary refers to non-speech recordings that were made of intake data during a study that recorded speech intake data (e.g. images, weight). Recordings are of a variety of data types and may be captured alongside speech to add extra detail or captured instead where speech would not be feasible. It does not refer to recordings made of non-intake data, such as physical activity and heart rate data from a wearable device or phone usage statistics. It is also separate from comparative recordings described in study design, where comparative recordings are used to evaluate speech, complementary recordings add to or assist with speech.

4) STRUCTURE

The format that a food item is described in can be structured or unstructured. Unstructured recordings do not require the participant to present their description in any particular format. Instead they generally present all details in a single input (e.g. ''Today for lunch I had a ham sandwich and a large apple''). This may also be referred to as natural language, and the pattern of speech can vary greatly between different individuals even whilst describing identical foods. Studies may specify the information that records should contain, such as quantity and brand name, but if the order of these items is unspecified it is unstructured data. Structured data are where participants must record data in a certain order or matching a certain format. This is often done by breaking the task into smaller portions, each requiring a discrete section of the overall description (e.g. ''a ham sandwich'' then ''apple'', ''large'').

C. DATA PROCESSING

The basic process of converting speech records to dietary data consists of the four general steps of (1) transcription, (2) natural language processing, (3) matching, and (4) quantification. The exact method used in each study varies greatly, yet distinguishing these four general steps allows us to make the inputs and outputs of the different studies more comparable.

1) TRANSCRIPTION

Transcription is the process of converting audio of speech records to text, often to facilitate further processing. Traditionally this has been manually done by researchers, but recent technologies allow for automation by a computer service. The text should represent the audio as accurately as possible, without modifying the structure or content.

2) NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is one method of automatically identifying information contained in raw unstructured text. There is a variety of methods of performing NLP, but the common goal is 'tagging' each word with properties obtained from the text, such as subject, descriptors and groupings. Modern NLP tools use machine learning models to operate. This model is a trained set of rules that informs the system on how to make decisions. A recent development in machine learning is deep learning, which requires less explicit training.

3) MATCHING

When a food item is extracted from the description, it must be matched to an item in a food composition database (FCD) to determine nutrient content. FCDs generally consist of thousands of items. Common items in FCDs will often have several variations to account for different offerings in the food supply and cooking or preparation methods (e.g. 'bread, white'; 'bread, wholegrain'; 'bread, white, toasted'. A particular food name may also appear in many different foods (e.g. 'cheese, cheddar', 'sandwich, filled with cheese').

4) QUANTIFICATION

Once the food item is identified, it is usually in the form of nutrient content per 100g. To find the actual consumed

nutrients the amount of the food must be estimated or calculated as a weight. To facilitate quantification in some FCDs common quantities specific to the population and/or food supply are included as weights, for example 'thick slice' for bread, 'medium' for an apple, 'can' for soft drink. If a density or specific gravity is available for a food item, weight can be calculated using a volume measure of common household measures such as cups or tablespoons. The purpose of this step is to identify the quantity in the description of intake data collected and match this to the appropriate food item.

D. EVALUATION

In the context of this review, only evaluations on a function of speech are included. Evaluations on factors that the use of speech as an input method do not affect are excluded, such as whether notifications aid participants in remembering to record. The impact of speech on three different factors was evaluated, the user acceptance, intake accuracy and processing accuracy.

1) USER ACCEPTANCE

This evaluation broadly covers the user experience with recording their intake as speech. Only the user acceptance of speech components is included, such as in apps that allow multiple input methods acceptance of other inputs is not reported. Other inputs may be noted where a comparison to speech is reported.

2) REPORTING ACCURACY

The accuracy of reporting intake refers to an evaluation of the accuracy of food intake collected through speech compared to reference method(s). This is a measure of how accurately a participant can record their intake with the given system; for example, comparing estimated energy intake derived from a participant's food descriptions to that participant's total energy expenditure as measured by an objective measure such as Doubly Labelled Water [5].

3) PROCESSING ACCURACY

Processing accuracy is a measure of how effectively a system can automate the process of analysing speech records, in the manner presented in section C. Processing. It differs from intake accuracy as it does not report a comparison to the actual intake of a participant.

V. RESULTS

The search of six databases retrieved 137 unique records. Following title and abstract screening, 92 records were excluded. The remaining 45 records proceeded to full-text review, with 28 records excluded. Forward and backward searches retrieved and included an additional four articles. The total number of included studies in this review was 21.

This section presents the extracted results of the literature in the format of the conceptual framework, illustrated in Figure 4. The structure of the conceptual framework resembles the method of performing dietary assessment on spoken descriptions as observed in the retrieved literature. After the study is defined, the data must first be collected by recording. The collected recordings are then generally processed, to extract relevant information and convert it to more usable data. Evaluation can then be performed on the results of the study. Not all studies contain processing or analysis. Some studies focus only on part of the process, such as performing NLP on recorded speech or evaluating user acceptance of recording methods. The conceptual framework allows these studies to be compared to others that contain different steps where they overlap. A detailed overview of the results for the four main components of the framework is provided in the supplemental materials S1-S5.

A. STUDY DESIGN

1) PARTICIPANTS

Only participants whose speech was recorded were included in this report. A total of 687 participants were included across all 21 studies. The median sample size was 21 (range 2 to 94). Age of participants was reported in 14 studies and ranged from 6.5 to 90 years. Six studies focused on elderly participants 60 years and older [16], [19]–[22], with one of these extended to also include middle-aged adults [5]. One study included only children (range 6.5 - 11.6 years) [23]. The remainder did not specify an explicit age range but stated recruiting adult participants aged 18 years and above. The median age range amongst all studies was 22.5 years.

Fifteen studies reported the gender of participants totalling 358 female and 172 male participants. A mix of female and male participants was included in 12 studies [4], [15], [16], [19]–[23], [27], [29], [30], [41], females only by two studies [5], [13] and males only by one study [24]. The median ratio of female to male participants was 16 to 12 (range 1-73 to 1-34).

Two studies ensured a population with varying literacy levels was used [4], [13]. Three studies only included participants with a chronic health condition, either having diabetes ($n = 1$) [29] or requiring dialysis ($n = 2$) [4], [13]. One study included a variety of accents [28].

2) SETTING

The setting in which recordings took place fell into two categories: free living, where the participant has full control over food selection ($n = 17$) [4], [5], [8], [13]–[15], [19]–[21], [23]–[25], [27], [29], [30], [36], [41] and controlled, where the researchers have that control ($n = 4$). Where the recording was controlled researchers either: created a script for participants to read ($n = 2$) [26], [28]; pre-made the meal participants were asked to describe ($n = 1$) [22]; or provided a limited set of items ($n = 1$) [16].

3) COMPARATOR

Comparative recordings were taken to assess accuracy of speech recordings by nine studies. Five studies recorded the participants energy expenditure (EE) through doubly labelled

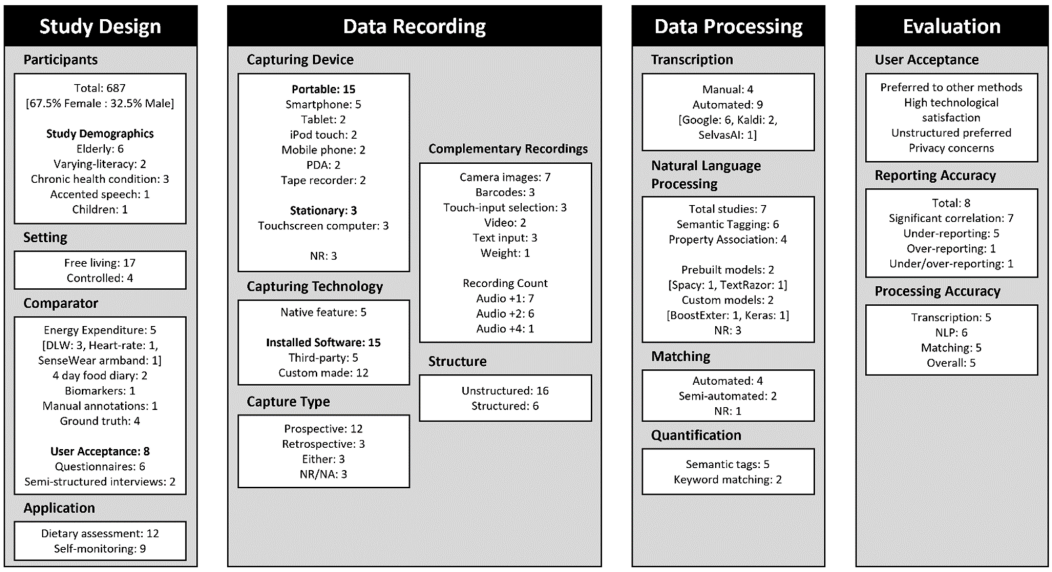


FIGURE 4. Summary of results of the systematic literature review.

water (n = 3) [5], [23], [29], a heart-rate monitor (n = 1) [24] or a SenseWear armband (n = 1) [41]. Two studies asked participants to complete a 4-day food diary at a later date, which would then be analysed by a dietitian [20], [21]. One study collected participants' urine and blood biomarkers to compare to protein and vitamin C intake [20].

Four studies recorded some form of ground truth data for comparison with participant collected data. Three of these were studies in a controlled setting, so food identity and quantity were recorded independently by researchers and thus known [22], [26], [28]. The last study had participants freely pick food from a cafeteria that was uncontrolled, but had researchers record the exact items and weight before consumption [30].

User acceptance was assessed by eight studies, primarily through questionnaires (n = 6) [4], [13], [14], [22], [29], [36] and semi-structured interviews (n = 2) [16], [27]. Two studies had participants use separate methods to record intake for comparison, namely traditional weighed food records (n = 1) [29], a commercially available app MyFitnessPal (n = 1) [14], and image-based recording (n = 1) [14]. Two studies developed multiple recording methods to contrast with each other [4], [22]. Two studies collected the time participants spent recording food descriptions [4], [22].

4) APPLICATION

Collected dietary data was used for dietary assessment (n = 12) [5], [19]–[25], [27], [29], [30], [41] or self-monitoring (n = 9) [4], [8], [13]–[16], [26], [28], [36]. Despite having less results, self-monitoring has been a more popular application since 2017.

Three studies applied speech to the context of managing chronic illness [4], [13], [29]. Monitoring dietary health for elderly users was likewise targeted by five studies [16], [19]–[22]. Obesity and the accompanying health risks were often mentioned as reasons for performing both dietary assessment and self-monitoring. Yet, no study was specifically designed to address obesity.

B. DATA RECORDING

1) CAPTURING DEVICE

The device used to capture participant speech was reported by 18 studies. Devices used were either portable (n = 15) or stationary (n = 3). Portable devices included: Microcassette tape recorders (n = 2) [5], [23]; Personal Digital Assistants (n = 2) [4], [13]; Mobile phones (n = 2) [25], [29]; touch-based iPods (n = 2) [24], [30]; Tablet computers (n = 2) [16], [22]; and Smartphones (n = 5) [14], [15], [27], [28], [41]. All three studies with a stationary device used a touch-screen computer [19]–[21].

Five of these seven devices utilised touchscreens (excluding tape recorders and pre-touchscreen mobile phones). All portable devices except the tablets were intended to be carried with the participant to be available at all mealtimes.

Of the three studies that did not report a capture device, two recorded speech read from a script and did not report the device used to do so [8], [26]. The third allowed for speech as data input but did not specify the device used to do so [36].

2) CAPTURING TECHNOLOGY

Some capturing devices provide a method of capturing speech as a native feature which five studies used without modification. The two studies using touch-based iPods utilised the native video record feature [24], [30]. Similarly, the two tape-recorder studies provided stock microcassette recorders [5], [23]. One mobile phone study used the phones native audio-recording capability [25].

The remaining 15 studies that reported a capture technology required some software to be used with the capture device hardware. This was either obtained from a third party

(n = 5) [4], [8], [14], [29], [41] or developed by the research team to fulfill the specific role (n = 12) [4], [13]–[16], [19]–[22], [27], [28], [36], with two studies requiring both custom and third-party software for participants to compare [4], [14]. Three of the third-party technologies were apps, FoodNow [41], MyFitnessPal [14] and Nutricam [29]. One study used Amazon Mechanical Turk to crowd-source the creation of food descriptions [8]. One study used an Integrated Voice Response System, a type of automated telephone system called Voxeo that participants could call to report foods [4]. The remaining 12 studies developed a custom tool specifically to collect food intake.

3) CAPTURE TYPE

Twelve studies only recorded speech descriptions in a prospective (at the time of consumption) manner. Three studies, all using the NANA system, only allowed for retrospective speech descriptions, forcing use of another method for prospective recordings [19]–[21]. Three studies allowed the user to record either prospective or retrospective as desired, although two of these expressed a heavy preference to participants to use prospective recording [4], [25]. Only one study did not specify when to record intake [14], but the impact that recording time has on accuracy and sentence structure was not explored. Three studies either did not report (n = 1) [36] or a recording type was not applicable as they generated a script (n = 2) [8], [26]. There were no studies that used speech recordings to capture recall-style descriptions such as 24hr recall or diet history.

4) COMPLEMENTARY RECORDINGS

Fourteen studies captured some form of complementary data in addition to speech. Types of data captured consisted of: camera images (n = 7) [5], [19]–[21], [27], [29], [41]; barcode scanning (n = 3) [4], [13], [16]; touch input (n = 3) [19]–[21]; video (n = 2) [24], [30]; text input (n = 3) [15], [27], [41]; and weight (n = 1) [27]. The number of data types recorded in addition to speech were: one additional (n = 7) [4], [5], [15], [16], [22], [24], [30]; two additional (n = 6) [13], [19]–[21], [29], [41]; and four additional (n = 1) [27]. Thirteen of the total 21 studies required speech as a compulsory recording [5], [8], [14], [16], [22]–[26], [28]–[30], [36]. Six of the 14 studies that captured additional complementary data had compulsory speech recording [5], [16], [22], [24], [29], [30].

5) STRUCTURE

Speech recordings of food descriptions were captured in structured (n = 6) [4], [16], [22], [26]–[28] and unstructured (n = 16) [4], [5], [8], [13]–[15], [19]–[21], [23]–[25], [29], [30], [36], [41] formats. One study [4] captured both formats for comparison. Three structured format studies required input in a specific order where the food item was identified before properties were recorded [4], [22], [27]. Two more used a script created by researchers to prompt speech [26], [28], and one required speech to match a certain format [16].

C. DATA PROCESSING

1) TRANSCRIPTION

Transcription was performed in 13 studies, either automated by a speech recognition engine (n = 9) [8], [14]–[16], [22], [26]–[28], [36] or manually performed by a human (n = 4) [4], [5], [25], [41].

For automated transcription, three different engines were used. Google's speech recognition as either a cloud resource or local package (n = 6) [14], [15], [22], [26]–[28] was the most popular, followed by an open-source solution Kaldi (n = 2) [8], [16], and lastly a commercially available package SelvasAI (n = 1) [36] which combines transcription and NLP tools.

All four studies performing transcription with humans used researchers to complete this task. Kaczkowski [5] allowed transcribers to view complementary images captured while transcribing, and Siek [4] had participants present while researchers transcribed descriptions.

Eight studies did not transcribe records, instead having researchers listen to speech records among other complementary data [13], [19]–[21], [23], [24], [29], [30].

2) NATURAL LANGUAGE PROCESSING

NLP was reported by seven studies [8], [14]–[16], [25], [26], [28]. Six of these studies utilised NLP to perform semantic tagging, where each word is attributed with a tag that denotes its usage in the text [8], [14], [15], [25], [26], [28]. This was used to detect the food items in descriptions, along with potential properties. Four studies furthered this process by performing property association, where each property tag is linked to the appropriate subject tag [8], [14], [15], [26]. This allowed for details like the quantity and cook method to be linked to the food item they apply to. The two studies that did not use property association used proximity to link identified properties to the correct food item [25], [28].

Four studies reported the tool used to perform NLP. These fell into two categories: prebuilt models and custom models. SpaCy [16] and TextRazor [28] use standard prebuilt models. BoostExter [25] and Keras [8] are toolkits that allow you to build models with your own data for greater customization. Three tools, SpaCy, TextRazor and BoostExter, use traditional machine learning models. Keras uses deep learning models. One study, Korpusik [8], compared two methods of semantic tagging. Semantic tagging of spoken data was found to be slightly more accurate and easier to implement using a deep learning Convolutional Neural Network, as opposed to a machine learning Conditional Random Field.

3) MATCHING

Seven studies matched their identified food items to a single item from a FCD [4], [8], [14], [15], [22], [26], [27]. Three studies reported using a combination of exact and approximate matching [14], [15], [26], where exact matching was attempted first and approximate used as a fallback. Two of these studies reported using edit-distance (Levenshtein

algorithm) for approximate matching, where the number of operations needed to convert one string to another is calculated [14], [15]. The third did not specify an approximate matching method [26].

Two studies had the user assist in matching, allowing them to select a single food item from a shortened list of possible matches [22], [27]. They used the input text to reduce the list from thousands of results to a handful of relevant matches. Liu [22] limits the number of displayed items to the top five. Neither paper reports the method of identifying matches in detail. One study had a teleprompt system with a small database of common items that the system would match to the participant's description [4]. One did not report the matching method [8].

4) QUANTIFICATION

Six studies reported extracting a quantification, using either the semantic tags obtained in the NLP step (n = 5) [8], [14], [15], [25], [26] or keyword matching (n = 2) [14], [28]. Keyword matching was done by looking for any numerical data (n = 1) [28] or looking for common quantity keywords when semantic tagging failed (n = 1) [14].

All six studies that extracted a quantification reported matching it to the appropriate food item. This was determined by: distance, where quantity was associated with the closest food item (n = 2) [25], [28]; and property association, where quantity tag was associated with the food by NLP (n = 4) [8], [14], [15], [26]. One study, Korpusik [8], compared two methods for linking the quantity: segmented, where a 'noun chunk' of the text around the item is identified in the processing and all tags in it linked; and property association, using a custom model to find probability of a property being linked to each food. No clear advantage was demonstrated by either.

D. EVALUATION

1) USER ACCEPTANCE

Eight studies evaluated the user experience of recording speech descriptions of their meals. Evaluation was obtained through questionnaires (n = 6) [4], [13], [14], [22], [29], [36], or semi-structured interviews (n = 2) [16], [27]. Two studies noted a specific questionnaire, namely the System Usability Scale [22] and the Questionnaire for User Interface Satisfaction [4].

In the three studies that compared speech to another method, speech was found to be the preferred option [4], [14], [29]. Speech was described as being easy, working well and with high technical satisfaction. One study compared speech-only to speech-and-buttons input, finding that adding buttons roughly doubled input time with similar accuracy [22].

In a study where unstructured and structured recordings were compared, unstructured was preferred by participants and structured descriptions were found to require significant training for varying-literacy users [4].

Privacy was noted as a concern by three studies, in regard to sending voice to large companies for transcription (n = 1) [16] or describing foods out loud in public (n = 2) [14], [27]. Both studies evaluating privacy concerns of speaking in public found it made participants uncomfortable and they would avoid it in public but found it useful in private.

2) REPORTING ACCURACY

Eight studies evaluated the accuracy of calculated intake obtained from speech descriptions. The intake was compared to: estimated energy expenditure (n = 5) [5], [23], [24], [29], [41]; 4-day food diary (n = 2) [20], [21]; ground truth weighing (n = 1) [30]; and protein and vitamin C biomarkers (n = 1) [20]. Seven of these studies described the speech recording method as feasible, stating an accuracy comparable or favourable to traditional methods [5], [20], [21], [24], [29], [30], [41]. The sole study that did not recommend the method recruited only child participants and found low adherence and data quality with this population. [23].

Under-reporting of speech recording was noted by four studies [5], [20], [29], [41]. The level of under-reporting was found to be similar to traditional methods such as 4-day food diaries and weighed food records. Over-reporting was found by a single study, which had dietitians estimate food sizes from images rather than text [30]. One study with child participants reported both over- and under-reporting [23].

One study performed three separate collection periods of seven days each over three months to assess accuracy over time and adherence with repeated usage, finding there was no significant change over time [19]. Two studies that recorded time taken to complete collection found the voice-only option significantly faster than options involving buttons [4], [22].

3) PROCESSING ACCURACY

The accuracy of each stage in the processing was often presented separately, so each could be independently evaluated. This is illustrated in Table 1. Where several accuracies were reported by a study for the same stage (e.g. where accuracy in high- and low-quality audio is compared [26]), the highest value is presented.

Three studies reported the accuracy of automated speech-to-text performed by Google (n = 2) [14], [26] or Kaldi (n = 1) [8]. All three calculated accuracy using word-error rate. Two Hezarjaribi papers tested multiple environments, finding accuracy could drop as low as 63.26% and 76.89% in areas with significant background noise [14], [26].

One study evaluated the accuracy of Google's speech-to-text algorithm, but did not report an accuracy, instead describing errors [28]. Common errors were found to be: unfamiliar food items to participants, poor recognition of accents due to English accent training data, and pauses in words creating separations.

One study implemented a training module where errors found in collected data were corrected daily and added to the training set, so that it incrementally learned user habits [15]. This significantly decreased the error rate for a participant

TABLE 1. Processing accuracy results as percentages.

Study	Scripted	STT	NLP	Matching	Overall
[25]	Unscripted	NA	90.0	NA	NA
[26]	Scripted	73.9	96.0	95.0	66.5
[8]	Unscripted	92.0	83.5	71.0	71.0
[28]	Unscripted	NR	NR	NR	NR
[14]	Scripted	82.0	NR	NR	97.7
[15]	Unscripted	NR	NR	NR	89.7

STT = Speech-to-Text, NLP = Natural Language Processing, NA = Stage was not executed, NR = Stage executed but accuracy not reported, Scripted = Descriptions were written by researchers before being read aloud

over several days and improved the starting accuracy of the next participant. The same study found that detecting missing quantity tags and prompting the user for input increased accuracy, but this would need further study.

VI. DISCUSSION

The aim of this review was to establish the current state of research on using speech recordings for dietary assessment. Speech recordings can be an effective means to overcome barriers to collecting intake data as recent advances in technology have considerably lowered the time and financial costs of managing the novel data. Recent research has focused on automated processing of speech data and shows promise but requires refinement and evaluation in more realistic scenarios.

A. SPEECH RECORDING

Food recording methods using speech were well-received among the evaluated populations, supporting the hypothesis that speech input can reduce burdens associated with recording and thereby improving adherence. While speech-based methods were preferred by participants to those without, it is important to note that only three studies compared speech to another method and hence such comparisons warrant further research attention. Also, when there was a choice between text input and voice input, text was often still selected. This is likely due to privacy concerns and feelings of discomfort when recording voice in public places [16]. As privacy was noted as a particular concern of users [16], [32] using speech recordings may introduce a new form of bias if data collection requires using speech in public venues. This potential bias requires further investigation to mitigate data loss.

It was found that speech recording allowed for a similarly accurate representation of actual intake as traditional methods. The primary identified benefit of using speech as an input method was the accessibility, as the literacy and complexity burdens are reduced [6], [13]. This allowed for usage by populations that would have difficulty with the traditional methods (e.g. users with low literacy).

Benefits of this input method are greatest when the speech recordings are unstructured [15]. Structured speech requires training, raising the complexity and literacy required and potentially introducing error when input does not match structure [4]. However, unstructured input runs the risk of not capturing sufficient detail for identification. Another issue identified was participants with low literacy having difficulty conveying detail about consumed foods (e.g. not being able to read the packaging [4]). In traditional methods led by an analyst, follow up questions would be asked when identification does not contain sufficient detail [11].

To allow for more accurate identification, images were often captured in conjunction with speech. Accurate portion size estimation requires training and often uses aids such as reference images and measures the user would not have access to (e.g. average unit size, density) [31]. Speech was found to be capable of containing more detail than images for mixed dishes (e.g. stews), or those where identity is unclear (e.g. juices) [14], but training on how to describe items was required for low-literacy populations [4]. As such, speech-image combinations were used in a third of the 21 studies [5], [19], [20], [21], [24], [27], [29] to ensure accurate identification and quantification, while not significantly increasing the burden.

With six studies specifically selecting older participants and a relatively high median age across all, the unique barriers experienced by older participants should influence study design [33]. The use of voice input has been suggested as a tool to assist older users with smartphone usage [34], but this must be formally evaluated. Additionally, the few studies that did not use a portable device specifically targeted elderly users, where the screen size is a consideration for usability [16], [34]. There was no study explicitly targeting adolescent or young adult participants. As age is known to affect smartphone adoption and usage accuracy [33], solutions may have different results when utilised by young participants. The sole study that did not find significant correlation between reported and actual intake can be explained by the unique participant demographic, self-reporting with speech is not a strong method of collecting intake from children (range 6.5 - 11.6 years) [23]. The age when speech becomes a feasible recording method requires further research.

Gender ratios are highly disparate, with almost twice as many female participants as male despite only two studies specifying only female participants. This is not addressed in reviewed literature but is a common trend in diet studies where female participants are noted to be more engaged and have higher compliance [35]. The number of studies not reporting gender is concerning considering the difference in diet behaviours [35].

B. APPLICATION

While the reasons for the use of speech were consistent (natural way of communication, reduced burden for participants, advances in speech recognition), the application varied. Dietary assessment using speech was popular, as it allows for more accurate assessment for clinical assessment or research on population intake [23], [24]. Self-monitoring was the focus of more recent research, likely due to the adoption of smartphones and the improving capability to automate data extraction from speech. The use of speech in other dietary assessment methods such as 24hr recalls were not addressed by any studies. The potential of new

technologies to reduce time and cost burdens associated with performing dietary intake assessment studies could allow for more comprehensive nutrition research. Unstructured speech is often described as simple and fast [6], [15], [25]. This input method may positively affect reactivity bias caused by complex and slow recording in prospective methods such as written weighed food records. The capability of speech recording to overcome varying literacy barriers [4] may also allow for data collection in populations where it has been unfeasible such as Low and Low-Middle Income Countries [3]. This also benefits self-monitoring, particularly in studies where a user has a health condition or has been prescribed a diet as part of medical nutrition therapy [4], [29], [36]. Prospective recording does come with the concern of introducing reporting bias due to complexity and time cost [2], [11], but it is hoped that this would be alleviated by the use of speech recognition and processing [15].

Smartphones provide an effective vehicle for collecting intake, being portable and thus always on-hand, able to record input, and having the computing power to process it and potentially provide feedback [41]. As such, all studies collecting actual intake data since the advent of smartphones have developed a custom application to handle this process [14], [15], [24], [27], [28]. The custom-built applications are best described as prototypes, with only a couple of features meant to be evaluated. They are also developed primarily for research data collection, providing little-to-no communication to the user in terms of feedback or health information. As the literature advances and the use of speech is improved, these applications will likely evolve to become closer to consumer-grade products. The design of such offerings is not trivial and will require significant further study. As seen in one computing study that contrasted two user interfaces [22], system complexity does impact the usability and therefore data accuracy. While speech may be validated as an input method, when novel features are added that significantly increase complexity the system should be evaluated independently in a nutrition context [6], [11]. Potential issues will only be exacerbated by the focus on accessibility, and factors like literacy must be considered if the full benefits are to be found [2], [4].

C. DISCIPLINE-SPECIFIC CHALLENGES

There are notable differences in the way that food items are represented in nutrition as compared to computing studies. Recent computing studies have aimed to minimise the involvement of human analysts or even replace them entirely, providing an automated solution that would allow a user to self-monitor their diet. When explaining the process of matching a descriptive string to the relevant nutritional information, several studies described ''exact'' matching [14], [15], [26]. However, it is important to note that it would be difficult for an untrained user to describe a food item that facilitates a perfect match to a comprehensive food composition database. An example given by a particular study [15] is the identification of a food item ''almonds''. It is matched to an item in the database ''almond''. Reviewing the database utilised by that study, there is no exact food name match for this item. The closest match is ''Nuts, almonds''. ''Almond'' does exactly match a subset of the database item name, but by this logic it also matches the incorrect item ''Snacks, granola bars, soft, almond''. From the perspectives of dietetics and nutrition, ''the name of a food is frequently insufficient for its unequivocal identification'' [38]. Factors such as the cooking method, branding, added nutrients, and preservation method contribute to the nutrient profile, and are represented in the food composition database [38]. If identification is performed by trained staff with the aid of complementary data the correct match can be inferred, but there is no mechanism explained to automate this process in reviewed studies. The challenge for research in this space is further compounded by the fact that there is not a country-specific FCD for all countries. Where a country does not have their own FCD, a database from another country/ies and/or regional FCD is used which may not reflect regional differences in the selection and processing of food. If a solution is to be applied internationally, the process of selecting the most appropriate food item must be explained.

D. ADVANCES IN MACHINE LEARNING

Automated speech transcription has driven recent research, with all studies since 2016, using a speech recognition engine to convert audio to text. The improvements in the speech recognition field can be seen in comparing transcription results in Hezarjaribi 2016 [26] and Hezarjaribi 2018 [14]. Almost identical experiments evaluate Google's transcription accuracy, with an 8.13% accuracy increase in well-recorded audio and a 13.63% increase on poor-quality audio after just two years of progress. The accuracy of this service is at a level where some studies did not feel it necessary to validate or report its accuracy. Most studies used this commercial service from Google. Commercial speech recognition models are built using data from a wide variety of domains. This comes at the drawback that descriptions unique to food may be difficult to translate (e.g. cornflakes, Massamun curry, congee), especially names originating in other languages [28]. Models constructed with domain specific data appeared to be more accurate [8], but this is a considerable investment. One function of machine learning is the capability to take a generalized model and refine it for a more specific task with significantly less data than training a new model entirely. In fact, the transcription service offered by Google popular in reviewed studies has recently begun offering this functionality [39] which would allow developers to add domain-specific data and address this shortcoming. This could benefit users with strong accents in a language or regions with specific terms (such as local food names), where a general model can be further specified to suit that population. This may mitigate the errors in transcription that previous research had attributed to accented speech [28]. As this technology is only just emerging, it has not been utilised by any study so far.

There was a noticeable trend that speech is gaining further recognition as technology has improved to process it. Earlier studies treated speech as a secondary input, yet the reduced processing burden has allowed recent studies to treat it as the sole or primary input. However, validated nutrition studies used speech primarily for identification, with complementary recordings taken for quantification [5], [24], [29], [30]. Particularly in terms of self-monitoring, portion size estimated by a user will not be accurate without the assistance of external tools. There is ongoing research into the automation of portion sizes in meals from images that could complement automated identification from speech. [40].

E. STRENGTHS AND LIMITATIONS

This review is the first comprehensive overview of the use of speech recording in the collection of dietary intake data. It illustrates the goals, benefits, and challenges with this input method, and the tools and procedures used by emerging literature to automate analysis. With authors from both computing and nutrition domains, the review synthesises the results from interdisciplinary literature to structure the insights of the literature in this context along a conceptual framework. We hope this framework will inform future research on the goals and process of recording intake for dietary studies.

Only texts published in English could be reviewed. While none were excluded by this constraint, there may be relevant literature unavailable through the databases utilised. To limit search-string length the wildcard operator was used to represent similar terms (e.g. diet* diary = diet record AND dietary record). As such the search string will not execute as presented if the wildcard operator is not supported by the database. Overall, it should be noted that there is a significant lack of research on speech (as a mode of data collection) in dietary assessment compared to other, more widely-used methods such as image and text. Finally, conclusions regarding the usability of the approach need to be taken with caution as only eight studies assess usability [4], [13], [14], [16], [22], [27], [29], [36], and only three of these compared speech recordings with another method [4], [14], [29].

VII. CONCLUSION

Reviewed literature confirms that speech input is an effective method of collecting dietary intake. It exhibits high user acceptance and significantly decreases barriers to recording, such as literacy level, cognitive load, and time [4], [13], [15], [20], [25]. The recorded intake has high correlation with actual intake, with similar accuracy to traditional methods [5], [21]. Speech recognition technologies are driving recent research, with the ability to automatically transcribe and process speech data [8] reducing the burdens of dietary assessment. While attempts to automate the process of extracting dietary data from speech records have yielded promising results, the technology is in its infancy and has not yet been developed or evaluated more broadly for free-living usage.

Further research is needed to consider the real-world complexities of the nutrition context, utilising the automation demonstrated by computing studies to help address challenges in dietary assessment and monitoring. A potential approach is using speech to gather dietary data in populations where it has previously been prohibitive, such as countries with low literacy rates [3]. Another avenue is the development of a user-centric application, using speech to improve adherence for dietary self-monitoring. The design of such an application should consider the populations that would benefit most from this and adapt to suit their unique needs. The use of other novel technologies such as digital personal assistants allowing for interactive speech interfaces may also be beneficial. Evaluating the consumer acceptability of an application employing these elements would greatly advance the literature.

ACKNOWLEDGMENT

This work was support in part by the Bill & Melinda Gates Foundation OPP1171389. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. Connor Dodd was additionally supported by an Australian Government Research Training (RTP) Scholarship.

REFERENCES

[1] A. Afshin, P. J. Sur, K. A. Fay, L. Cornaby, G. Ferrara, J. S. Salama, E. C. Mullany, K. H. Abate, C. Abbafati, Z. Abebe, and M. Afarideh, "Health effects of dietary risks in 195 countries, 1990–2017: A systematic analysis for the global burden of disease study 2017," *Lancet*, vol. 393, no. 10184, pp. 1958–1972, 2019.

[2] F. Cordeiro, D. A. Epstein, E. Thomaz, E. Bales, A. K. Jagannathan, G. D. Abowd, and J. Fogarty, "Barriers and negative nudges: Exploring challenges in food journaling," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 1159–1162.

[3] W. Bell, B. A. Colaiezzi, C. S. Prata, and J. C. Coates, "Scaling up dietary data for decision-making in low-income countries: New technological frontiers," *Adv. Nutrition, Int. Rev. J.*, vol. 8, no. 6, pp. 916–932, Nov. 2017.

[4] K. A. Siek, "Evaluation of two mobile nutrition tracking applications for chronically ill populations with low literacy skills," in *Mobile Health Solutions for Biomedical Applications*. Hershey, PA, USA: IGI Global, 2009, pp. 1–23.

[5] C. H. Kaczkowski, P. J. H. Jones, J. Feng, and H. S. Bayley, "Four-day multimedia diet records underestimate energy needs in middle-aged and elderly women as determined by doubly-labeled water," *J. Nutrition*, vol. 130, no. 4, pp. 802–805, Apr. 2000.

[6] P. R. Cohen and S. L. Oviatt, "The role of voice input for human-machine communication," *Proc. Nat. Acad. Sci. USA*, vol. 92, no. 22, pp. 9921–9927, Oct. 1995.

[7] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki, and D. Mitsouras, "Natural language processing technologies in radiology research and clinical applications," *Radiographics*, vol. 36, no. 1, pp. 176–191, Jan. 2016.

[8] M. Korpusik and J. Glass, "Spoken language understanding for a nutrition dialogue system," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1450–1461, Jul. 2017.

[9] *Global status report on noncommunicable diseases 2010*. World Health Org., Geneva, Switzerland, 2011.

[10] K. A. Cradock, G. ÓLaighin, F. M. Finucane, R. McKay, L. R. Quinlan, K. A. Martin Ginis, and H. L. Gainforth, "Diet behavior change techniques in type 2 diabetes: A systematic review and meta-analysis," *Diabetes Care*, vol. 40, no. 12, pp. 1800–1810, Dec. 2017.

[11] F. E. Thompson and A. F. Subar, *Dietary Assessment Methodology*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 5–48.

[12] N. Vuckovic, C. Ritenbaugh, D. L. Taren, and M. Tobar, "A qualitative study of participants' experiences with dietary assessment," *J. Amer. Dietetic Assoc.*, vol. 100, no. 9, pp. 1023–1028, Sep. 2000.

[13] K. Connelly, K. A. Siek, B. Chaudry, J. Jones, K. Astroth, and J. L. Welch, "An offline mobile nutrition monitoring intervention for varying-literacy patients receiving hemodialysis: A pilot study examining usage and usability," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. 5, pp. 705–712, Sep. 2012.

[14] N. Hezarjaribi, S. Mazrouee, and H. Ghasemzadeh, "Speech2Health: A mobile framework for monitoring dietary composition from spoken data," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 252–264, Jan. 2018.

[15] N. Hezarjaribi, S. Mazrouee, S. Hemati, N. S. Chaytor, M. Perrigue, and H. Ghasemzadeh, "Human-in-the-loop learning for personalized diet monitoring from unstructured mobile data," *ACM Trans. Interact. Intell. Syst.*, vol. 9, no. 4, pp. 1–24, Dec. 2019.

[16] A. Seiderer, H. Ritschel, and E. André, "Development of a privacy-by-design speech assistant providing nutrient information for German seniors," in *Proc. 6th EAI Int. Conf. Smart Objects Technol. Social Good*, Sep. 2020, pp. 114–119.

[17] R. F. Baumeister and M. R. Leary, "Writing narrative literature reviews," *Rev. Gen. Psychol.*, vol. 1, no. 3, pp. 311–320, Sep. 1997.

[18] N. H. Chowdhury, M. T. P. Adam, and G. Skinner, "The impact of time pressure on cybersecurity behaviour: A systematic literature review," *Behaviour Inf. Technol.*, vol. 38, no. 12, pp. 1290–1308, Dec. 2019.

[19] C. Moore, C. M. Timon, L. Maclean, F. Hwang, T. Smith, T. Adlam, A. J. Astell, and E. A. Williams, "Use of NANA, a novel method of dietary assessment, for the longitudinal capture of dietary intake in older adults," *Proc. Nutrition Soc.*, vol. 72, no. OCE4, p. E267, 2013.

[20] C. M. Timon, "The validation of a computer-based food record for older adults: The novel assessment of nutrition and ageing (NANA) method," *Brit. J. Nutrition*, vol. 113, no. 4, pp. 654–664, 2015.

[21] A. J. Astell, F. Hwang, L. J. E. Brown, C. Timon, L. M. Maclean, T. Smith, T. Adlam, H. Khadra, and E. A. Williams, "Validation of the NANA (novel assessment of nutrition and ageing) touch screen system for use at home by older adults," *Experim. Gerontol.*, vol. 60, pp. 100–107, Dec. 2014.

[22] Y.-C. Liu, C.-H. Chen, Y.-S. Lin, H.-Y. Chen, D. Irianti, T.-N. Jen, J.-Y. Yeh, and S. Y.-H. Chiu, "Design and usability evaluation of mobile voice-added food reporting for elderly people: Randomized controlled trial," *JMIR mHealth uHealth*, vol. 8, no. 9, Sep. 2020, Art. no. e20317.

[23] C. H. Lindquist, T. Cummings, and M. I. Goran, "Use of tape-recorded food records in assessing Children's dietary intake," *Obesity Res.*, vol. 8, no. 1, pp. 2–11, Jan. 2000.

[24] A. H. Robertson, C. Larivière, C. R. Leduc, Z. McGillis, T. Eger, A. Godwin, M. Larivière, and S. C. Dorman, "Novel tools in determining the physiological demands and nutritional practices of Ontario FireRangers during fire deployments," *PLoS ONE*, vol. 12, no. 1, Jan. 2017, Art. no. e0169390.

[25] R. Lacson and W. Long, "Natural language processing of spoken diet records (SDRs)," in *Proc. Annu. Symp. (AMIA)*, 2006, pp. 454–458.

[26] N. Hezarjaribi, C. A. Reynolds, D. T. Miller, N. Chaytor, and H. Ghasemzadeh, "S2NI: A mobile platform for nutrition monitoring from spoken data," in *Proc. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 1991–1994.

[27] A. Seiderer, S. Flutura, and E. André, "Development of a mobile multi-device nutrition logger," in *Proc. 2nd ACM SIGCHI Int. Workshop Multisensory Approaches Hum.-Food Interact.*, Nov. 2017, pp. 5–12.

[28] H. Oh, J. Nguyen, S. Soundararajan, and R. Jain, "Multimodal food journaling," in *Proc. 3rd Int. Workshop Multimedia Pers. Health Health Care.*, Seoul, South Korea, 2018, pp. 39–47.

[29] M. E. Rollo, S. Ash, P. Lyons-Wall, and A. Russell, "Trial of a mobile phone method for recording dietary intake in adults with type 2 diabetes: Evaluation and implications for future applications," *J. Telemedicine Telecare*, vol. 17, no. 6, pp. 318–323, Sep. 2011.

[30] E. Jago, A. P. Gauthier, A. Pegoraro, and S. C. Dorman, "An assessment of the validity of an audio-video method of food journaling for dietary quantity and quality," *J. Nutrition Metabolism*, vol. 2019, pp. 1–8, Mar. 2019.

[31] R. Z. Franco, R. Fallaize, J. A. Lovegrove, and F. Hwang, "Popular nutrition-related mobile apps: A feature assessment," *JMIR mHealth uHealth*, vol. 4, no. 3, p. e85, Aug. 2016.

[32] A. Seiderer and E. André, "Development of a multi-device nutrition logging prototype including a smartscale," in *Proc. Int. Conf. Digit. Health*, London, U.K., Jul. 2017, pp. 239–240.

[33] H. M. Mohadisdudis and N. M. Ali, "A study of smartphone usage and barriers among the elderly," in *Proc. 3rd Int. Conf. User Sci. Eng. (i-USEr)*, Sep. 2014, pp. 109–114.

[34] M. S. Dias, "Multimodal user interfaces to improve social integration of elderly and mobility impaired," in *Proc. 9th Int. Conf. Wearable Micro Nano Technol. Pers. Health*. Portugal: IOS Press, 2012.

[35] F. Babwah, S. Baksh, L. Blake, J. Cupid-Thuesday, I. Hosein, A. Sookhai, C. Poon-King, and G. Hutchinson, "The role of gender in compliance and attendance at an outpatient clinic for type 2 diabetes mellitus in trinidad," *Revista Panamericana de Salud Pública*, vol. 19, no. 2, pp. 79–84, Feb. 2006.

[36] K. Chung and J. Kim, "Activity-based nutrition management model for healthcare using similar group analysis," *Technol. Health Care*, vol. 27, no. 5, pp. 473–485, Sep. 2019.

[37] J. L. Welch, K. A. Siek, K. H. Connelly, K. S. Astroth, M. S. Mcmanus, L. Scott, S. Heo, and M. A. Kraus, "Merging health literacy with computer technology: Self-managing diet and fluid intake among adult hemodialysis patients," *Patient Educ. Counseling*, vol. 79, no. 2, pp. 192–198, May 2010.

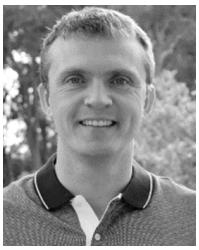
[38] H. Greenfield and D. A. Southgate, *Food Composition Data: Production, Management, and Use*. Rome, Italy: Food & Agriculture Org., 2003.

[39] Google. Improve transcription results with model adaptation 2021 [cited 2021; Available from:. [Online]. Available: https://cloud.google.com/speech-to-text/docs/adaptation-model.

[40] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2Calories: Towards an automated mobile vision food diary," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1233–1241.

[41] F. J. Pendergast, N. D. Ridgers, A. Worsley, and S. A. McNaughton, "Evaluation of a smartphone food diary application using objectively measured energy expenditure," *Int. J. Behav. Nutrition Phys. Activity*, vol. 14, no. 1, pp. 1–10, Dec. 2017.

CONNOR T. DODD received the B.Sc. degree (Hons.) in information technology from the University of Newcastle, Newcastle, Australia, where he is currently pursuing the Ph.D. degree in information technology. From 2018 to 2020, he was a Software Engineer at the University of Newcastle—as part of a project group aimed at improving nutritional knowledge and research in low and low-middle income countries with the use of new novel technologies. This project inspired his current research interest and the focus of his thesis.



MARC T. P. ADAM received the undergraduate degree in computer science from the University of Applied Sciences Würzburg, Germany, and the Ph.D. degree in information systems from the Karlsruhe Institute of Technology, Germany. He is an Associate Professor of computing and information technology with the University of Newcastle, Australia. His research focuses on the interplay of users' cognition and affect in human–computer interaction. He is a Founding Member of the Society for NeuroIS.

MEGAN E. ROLLO received the BAppSci, BHlthSci (Nutr&Diet), and Ph.D. degrees from the Queensland University of Technology, Australia. She is a Lecturer in nutrition and dietetics with the School of Population Health, Curtin University, Australia. Her research interests include technology-assisted dietary assessment and personalized behavioral nutrition interventions.